Development of a two-level full factorial model to analyse antenatal HIV data

MG Dhlamini

orcid.org 0000-0001-7105-2906

Dissertation accepted in fulfilment of the requirements for the degree *Masters of Science in Computer Science* at the North West University

Supervisor: Prof PD Pretorius

Co-supervisor: Dr W Sibanda

Graduation ceremony: October 2018

Student number: 22438114

DECLARATION

I Mantha Gabriela Dhlamini declare that

Development of a two-level full factorial model to analyse antenatal HIV data

Is my own work, and that all the sources I used or quoted have been indicated and acknowledged by means of complete references.

Signature: _____

Date: October 2018

ACKNOWLEDGEMENTS

The journey of research leading to the compilation of this dissertation has been a long and by no means simple one. The journey has led me down the path of self-discovery, and I would like to acknowledge the following people who made this expedition possible.

First and foremost, I would like to praise and honour God for his love and strength. Without him I am helpless.

Second, I would like to give special thanks to Prof. Philip Pretorius, my supervisor, for his input, guidance and patience. Thank you for not giving up on me.

Third, I would like to thank my co-supervisor for sharing his knowledge and data with me. Your expertise has really helped me to grow.

I would also like to give a special thanks to Prof. Babs for affording me this great opportunity that has changed my life.

Finally, I would like to thank Thabo Ramalitse for his support and assistance whenever I needed it.

Special thanks are due to my family: I would like to thank my sisters (Evelyn, Veronica and Jane) for their support and prayers, and above all my father and mother, Sello and Selina Dhlamini for believing in me and supporting me throughout my study. I dedicate this thesis to you, Mom and Dad.

> *Being confident of this, that he who began a good work in you will carry it on to completion until the day of Christ Jesus*
>
> Philippians 1: 6.

UNICUIQUE SUUM

ABSTRACT

The present study is based on antenatal HIV data collected annually by South Africa's National Department of Health (NDoH) from 2001 to 2010. The data was obtained by sampling pregnant women attending the clinic for antenatal care for the first time.

The main research questions of this study are as follows:

1. Is it possible to develop two-level full factorial models to analyse coded antenatal HIV data for each year?
2. Do the models remain the same over the years?

This study describes the development of two-level full factorial models to assist in analysing and understanding coded HIV antenatal sample data from 2001 to 2010.

The development of the two-level full factorial models was done by developing two-level full factorial matrices and using them to estimate HIV risk models. This was done by using one demographic variable at a time for each year, and using all the demographic variables for each year. ANOVA is used to analyse and interpret the data.

In this study regression analysis was also directly applied to HIV data without estimating full factorial matrices. The regression analysis was used in developing HIV risk models for all of the ten years.

Simple linear regression models were used to model time trends.

The study concludes with a description of the findings and a summary of the chapters. Future research possibilities are discussed and recommendations for research are made.

Key words: HIV risk models, coded antenatal HIV data, design of experiments, two-level full factorial models, regression analysis, linear probability models.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION

Numerous studies on the analysis of the Human Immunodeficiency Virus (HIV) and syphilis among pregnant women have been conducted and various statistical methods have been used. An example of such a study is *The sero-conversion rate of syphilis and HIV among pregnant women attending antenatal clinic in Tanzania* (Lawi et al., 2015b).

This was a cross-sectional, hospital-based study of pregnant women attending the Buganda Medical Centre (BMC). The serum samples were collected using a standardised data collection tool and analysed using STATA version 11 (Lawi et al., 2015b). The study concluded that re-screening is necessary after birth to ensure that HIV and syphilis were not missed in the first screening (Lawi et al., 2015b).

The year 2012 marked 30 years since the first incident of the Human Immunodeficiency Virus (HIV) was reported and 15 years since HIV treatment became a reality (NDoH, 2012). However, despite cost-effective treatment which has become available to the general public, HIV and syphilis infections are still common among pregnant women in the Sub-Saharan region of Africa (Lawi et al., 2015b).

In light of this, the South African National Department of Health (NDoH) introduced a new method to monitor the HIV epidemic on an annual basis since 1990, which was achieved by conducting annual nation-wide HIV and syphilis sero-prevalence surveys among pregnant women attending public-sector antenatal clinics (NDoH, 2012).

The use of data mining and statistical methods are extremely important to the understanding and analysis of how the behaviour of the HIV epidemic has changed over the years (Sibanda, 2013). This study therefore seeks to develop a two-level full factorial model to enable the HIV antenatal data to be analysed.

1.2 PROBLEM STATEMENT

Sibanda and Pretorius conducted a study in which a two-level fractional design was used to develop and optimise the combination of demographic characteristics that has

the greatest effect on the spread of HIV in South Africa. They concluded that the study was successful (Sibanda and Pretorius, 2011).

The HIV data collected at the antenatal clinics include demographics such as: the pregnant woman's age, level of education, gravidity (defined as the number of pregnancies the woman has had), parity (defined as the number of children the woman has), the age of the woman's partner as well as the pregnant woman's HIV and syphilis results (Sibanda and Pretorius, 2014).

Taking into consideration the literature that has been discussed, it is evident that a two-level full factorial model with the use of two-level full factorial models has not been used to analyse HIV antenatal data. Given that only a two-level fractional model was used, this study intends to fill the gap in the literature by developing a two-level full factorial model to analyse antenatal HIV data.

The research questions to this study are the following:

1. Is it possible to develop a two-level full factorial model to analyse antenatal HIV data?
2. Do the HIV risk models change or remain the same over time?

1.3 MOTIVATION OF THE STUDY

The first case of HIV in South Africa was reported in 1982, and the first AIDS death was recorded in 1985.

In 1990 the NDoH took it upon themselves to start the antenatal sentinel surveillance programme to monitor the prevalence of HIV at national, provincial and district level. The antenatal sentinel surveillance data is HIV data collected on the basis of a blood survey conducted on pregnant women visiting antenatal clinics for their first check-up throughout the Republic of South Africa (NDoH, 2012).

Over time different mathematical and statistical methods or models were used with the aim of understanding the changes in the behaviour of the HIV epidemic.

The government of the Republic of South Africa has also over the years taken into consideration the various demographic characteristics of pregnant women with the intent of understanding factors that could contribute to the risk of HIV.

The data used in this study is coded because that was the only data made available to the researcher for the purpose of this study. The coded levels of the pregnant woman's demographic characteristics are defined as follow:

a. *Pregnant woman's age*

There were two groups of pregnant women from which data was collected, firstly pregnant women of ages 24 years and younger denoted by (-1), and pregnant women of ages 25 years and old denoted by (1).

b. Pregnant women's partners' age

The pregnant women's partners' age was also captured, and two age groups were formed: partners of ages 28 years and younger denoted by (-1), and partners of ages 29 years and older denoted by (1).

c. Pregnant women's gravidity

Gravidity is the number of times the pregnant woman has been pregnant before. Parity was also grouped into two categories, namely pregnant women their first pregnancy denoted by (-1), and pregnant women who had one or more pregnancies before denoted by (1).

d. Pregnant women's parity

Parity describes the number of children these currently pregnant women had. Similarly parity was grouped into two categories, pregnant women with no children denoted by (-1), and pregnant women who already had one or more children denoted by (1).

e. Pregnant women's level of education

The level of education of the pregnant women attending antenatal care for the first time was also captured, and was also placed into two categories, namely pregnant women with primary to no education denoted by (-1), and pregnant women with secondary to tertiary education denoted by (1).

f. Pregnant women's syphilis status

Syphilis is one of the leading contributors to the risk of HIV, and therefore the pregnant women's syphilis status was also recorded and placed into two categories: pregnant

women who tested negative for syphilis denoted by (-1), and pregnant women who tested positive for syphilis denoted by (1).

The main motivation for this study is to develop two-level full factorial models to analyse antenatal HIV data with the aim of understanding the effect of the demographic characteristics on the risk of HIV.

1.4 OBJECTIVES OF THE STUDY

The purpose of the study was to develop two-level full factorial models using multivariate analysis with the aim of considering all possible combinations of the pregnant woman's demographics. The analysis of the pregnant women was conducted on ten years' worth of HIV antenatal data (2001–2010) with the aim of understanding the differential effects of the demographic characteristics of the pregnant woman on the risk of HIV infection.

The study used coded data because it was the only data set made available to the researcher. The actual data was not made available.

The study makes use of a two-level full factorial model because there are two levels to each of the demographic characteristics of the pregnant women.

Pregnant women's demographic characteristics	**Level -1**	**Level 1**
Mother's age	Ages 13-25	Ages 26-40
Father's age	Ages 13-25	Ages 26-60
Gravidity	First pregnancy	One or more pregnancies
Parity	First child	One or more children
Education	Primary school to no education	Secondary to tertiary education
Syphilis status	Syphilis negative	Syphilis positive

The following objectives were formulated for the study:

1.4.1 Primary objective

The primary objective of this study was to develop two-level full factorial models to analyse antenatal HIV data on an annual basis.

1.4.2 Secondary objective

Evaluate whether the two-level full factorial models remained stationary over a 10-year period from 2001 to 2010.

1.4.3 Theoretical objectives

To achieve the primary objective, the following theoretical objectives were formulated:

A. Research the literature to gain a better understanding of design of experiments methodology.
B. Research the literature to gain a better understanding of two-level full factorial analysis.
C. Research the literature to gain a better understanding of data analysis.

1.4.4 Empirical objectives

In accordance with the primary objective of the study, the following empirical objectives were formulated:

A. Develop two-level full factorial models for the analysis of antenatal HIV data.
B. Evaluate whether the two-level full factorial models remain stationary over the ten-year period from 2001 to 2010.

1.5 RESEARCH DESIGN AND METHODOLOGY

This section describes the methodology chosen to conduct this research. It consists of positivism as a research paradigm, and the use of design of experiments focusing on two-level full factorial and multivariate analysis.

1.5.1 Literature study

According to Guba and Lincoln (1994), a paradigm defines how one views the world and everything that surrounds these views. There are three philosophical aspects, which Scotland (2012) identifies as: ontology, epistemology and methodology. Ontology is defined as the study of being. Epistemology is the study of how knowledge is created, acquired and communicated, and methodology is defined by Saunders et al. (2009) as the study of the manner in which research should be conducted.

Scotland (2012) identifies three types of research paradigms, namely positivist, interpretivists and social constructionist research paradigms.

1.5.2 Social constructionist research paradigm

Constructionists or critical researchers posit that social reality has always been present in the form of history, and is produced and reproduced by people (Aliyu et al., 2014). Wahyuni (2012) states that researchers who use the social constructionist paradigm are part of the research, meaning that they cannot be separated from the truth and are therefore subjective.

1.5.3 Interpretivism research paradigm

According to Aliyu et al. (2014), interpretivists posit that there are multiple methods of acquiring knowledge and that there is not just a single worldwide or universal truth. Research in this paradigm is conducted through the use of case studies, field experiments, exploratory analysis and qualitative analysis, and the research is directed at understanding the world or the truth from the individual's perspective (Scotland, 2012).

Individual philosophies are explained and understood through interaction between researcher and participants (Guba and Lincoln, 1994), state that interpretivists believe that knowledge and truth are discovered by interacting with the world and being conscious of one's surroundings (Scotland, 2012).

1.5.4 Positivist research paradigm

Krauss (2005) states that positivists' core argument is that the social world exists externally from the researcher. Positivists are concerned with attempting to identify causes that affect outcome (Scotland, 2012), and they believe that knowledge is acquired through the experience of the senses and can be attained through observations and experiments (Noor, 2008). The reality is observed and data is collected using senses

Positivism focuses on the gathering of quantitative data which is analysed by the use of statistical methods, with some focus on the relationship between the variables (Aliyu et al., 2014). Quantitative data is most often used in positivist studies (Saunders et al., 2009).

1.5.5 Design Science Research paradigm

There is a fourth research paradigm that has been introduced known as the Design Science Research paradigm (DSR) which Vaishnavi and Kuechler (2004) explain are a paradigm that introduces the development of artefacts to solve problems.

According to Peffers et al. (2007), DSR is a process of carefully designing artefacts to find solutions to challenges or problems, to contribute to research, to evaluate the designed artefacts and to communicate the results. Hevner et al. (2004) state that through the creation of new and innovative artefacts, DSR seeks to broaden the boundaries of human organisational capabilities.

Table 1.1 summarises the different research paradigms as well as their philosophical assumptions (Creswell, 2013) (Vaishnavi and Kuechler, 2004) and (Wahyuni, 2012).

Table 1-1: Research philosophical aspects

Research paradigm	**Ontology**	**Epistemology**	**Methodology**	**Axiology**
Positivist	Determination, reductionism, empirical observation and measurement, theory verification	Researcher is external, objective and independent of social factors	Experimental, quantitative, hypothesis testing	Truth Predictions
Interpretivist	Socially constructed, subjective, may change and has multiple realities	Observer is subjective	Interactional Qualitative	Researcher is part of study
Constructionist Critical social theory	Socially constructed reality	Suspicious, political Observer	Textual analysis	Value-bond Researcher's values affect the research

		constructs truth		
Design Science Research	Multiple, contextually situated realities	Knowing through doing	Developmental Impact analysis of artefact on composite system	Control Creation Understanding

The present study is positioned within the positivism research paradigm as it supports knowledge through survey sampling and the use of quantitative data. This study made use of antenatal HIV data and applied the design of experiments focusing on the development of two level full factorial models in the analysis of the antenatal HIV data.

1.5.6 Design of Experiments methodology

Design of Experiments (DOE) is a method that was invented by Ronald A. Fisher in 1920, and although it was initially developed for the agricultural sector, it has been successfully used by the military and in various industries. DOE is a method in which a sequence of tests are conducted, to which meaningful changes are made to the input variables of a system or a process and the effect on the response variables are measured (Telford, 2007).

A factorial design is a method used in DOE which Morris explains as a factorial treatment structure where the effect of many different factors or treatments are tested by varying them simultaneously (Morris, 2011). The use of a full-factorial design requires that an experimental run be performed with all combinations of each factor level (JMP, 2014).

However, this study is not based on experimental runs, but on available sample data. This means that the analysis is of the combinations that are available in the sample of each year.

See Appendix B: there are missing values for each of the years, and are therefore full factorial models with missing values.

1.6 EMPIRICAL STUDY

This study used HIV data that was collected by the South African National Department of Health (NDoH) during their annual national antenatal sero-prevalence survey conducted among pregnant women attending public-sector clinics for the first time.

The national annual antenatal sero-prevalence survey is conducted yearly during the month of October.

The empirical section that follows describes how the South African National Department of Health (NDoH) collected the data.

1.6.1 Target population

The NDoH's HIV and syphilis prevalence survey included pregnant women attending antenatal care at public clinics for their first appointment during their current pregnancy (NDoH, 2012).

1.6.2 Sampling frame

The sampling frame that was used by the NDoH comprised pregnant women attending antenatal care in nine provinces and 52 health districts (NDoH, 2012).

1.6.3 Sampling method

The National Department of Health used two different criteria to select the population that were to be included in the survey (NDoH, 2012), namely the inclusion criteria and the exclusion criteria.

a) **Inclusion criteria:** All pregnant women attending antenatal clinics for the first time during their current pregnancy were eligible for inclusion.
b) **Exclusion criteria:** Pregnant women who had previously visited antenatal clinics during their current pregnancy during the survey period were excluded – this was done to avoid redundancy in the data.

1.6.4 Sample size

There were 218 843 thousand pregnant women that the NDoH included in the survey in the period 2001 to 2010 (Sibanda and Pretorius, 2014).

1.6.5 Data collection method

The NDoH used surveys as their data collection method (NDoH, 2012).

1.7 STUDY LAYOUT

As mentioned, this study used available antenatal HIV sample data which was collected by the NDoH and made available to Dr Wilbert Sibanda for research purposes.

Chapter 1 gives the introduction to the study and Chapter 2 discusses the methodology.

Chapter 3 is a literature review that discusses statistical methods used in the study.

Chapter 4 takes a closer look at the pregnant woman's demographic characteristics, with the use of linear models, to determine if there are trends with the data.

Chapter 5 introduces the development of the two-level full factorial modes. The chapter also answers the two research questions in Chapter One.

Chapter 6 gives a summary of the entire study. It discusses the findings and gives conclusions of the findings and future research recommendations.

1.8 ETHICAL CONSIDERATIONS

The data has no identifiers and therefore no ethical considerations were required.

1.9 CHAPTER CLASSIFICATION

This section provides an overview of how the chapters are arranged and the concepts that are discussed in each chapter.

Chapter 1 Introduction: This chapter presents the introduction, problem statement and objectives of this research.

Chapter 2 Research Design and Methodology: This chapter provides more detail about the positivism research paradigm and the Design of Experiments methodology.

Chapter 3 Statistical Methods: This chapter provides literature on the statistical methods used in the study.

Chapter 4 Data Analysis: This chapter provides data analysis of the demographical characteristics of the pregnant women over time, with the use of linear models.

Chapter 5 Development of a Two-Level Full Factorial Model: This chapter shows the development of the two-level full factorial matrix and uses Anova to assist in the analysis of more than two variables, and also answers the question as to whether it was possible to develop a two-level full factorial model for the analysis of antenatal HIV data. Finally, it analyses and evaluates whether the model remained stationary over the years.

Chapter 5 Conclusions and Recommendations: This chapter concludes the study. It contains lessons learned, challenges encountered, as well as future opportunities and recommendations.

1.10 CHAPTER CONCLUSION

The objective of this chapter was to introduce the study and provide a study layout. This was achieved by introducing the problem statement and questions asked by the study, as well as by describing the objectives of the study and finally presenting the chapter classification.

Chapter 2 discusses the existing literature based on the research methodology.

CHAPTER 2: RESEARCH DESIGN AND METHODOLOGY

2.1 INTRODUCTION

In Chapter 1 the study objective was discussed which was to develop two-level full factorial models to analyse antenatal HIV data. A brief description of the research methodology and the research philosophies was given, which will be further discussed in this chapter.

As mentioned above, the primary objective of this study was to develop two-level full factorial models to analyse antenatal HIV data. To achieve this, a search of the literature on research methodology and Design of Experiments (DOE) pertaining to a two-level full factorial model was first required.

The two-level full factorial model is widely used mainly as it is easy to design, efficient to run and is also full of information that can be analysed (Boon and Mariatti, 2014). A full factorial model takes into consideration every combination of the factors in the experiment. For example, if we have k factors, each at two levels, then the full factorial consists of 2x2x...x2 = 2^k experimental runs (Boon and Mariatti, 2014).

All the factors considered in this study are each at two levels, hence the use of a two-level full factorial model. *The data used in this study was coded as it was the only data made available to the researcher.*

The term research is used to describe a logical and systematic manner of uncovering new and useful information on a specific subject. It enables the researcher to investigate new and innovative ways of solving problems and uncovering hidden truths (Rajasekar et al., 2013). The distinction between a method and methodology is often confused, and according to Rajasekar et al. (2013). The difference between the two is that a method consists of the various techniques, schemes and algorithms that are used in research, for example the statistical methods used, whereas a methodology refers to how research is to be conducted (Saunders et al., 2009).

The objective of this chapter is to demonstrate an understanding of the research methodology and how it contributes to the development of this study. This chapter also includes a discussion of research philosophies, paradigms and methods in general and also literature on Design of Experiments methodology.

The chapter is divided into the following sections: research philosophy (Section 2.2), research paradigms (Section 2.3), and research approaches (Section 2.4), Design of Experiments (Section 2.5), data collection (Section 2.6) and the conclusion (Section 2.7).

Saunders *et al*. (2009) explain the research approach using the comparison of an onion as shown in Figure 2.1, where the outer layers describe the different philosophies and paradigms that are applied in research. In the present study positivism is the research philosophy. The inner layers of the onion represent the strategy which will be used in the research, the choices and time horizon after which the researcher can move to the data collection and analysis part of the research (Kulatunga et al., 2007).

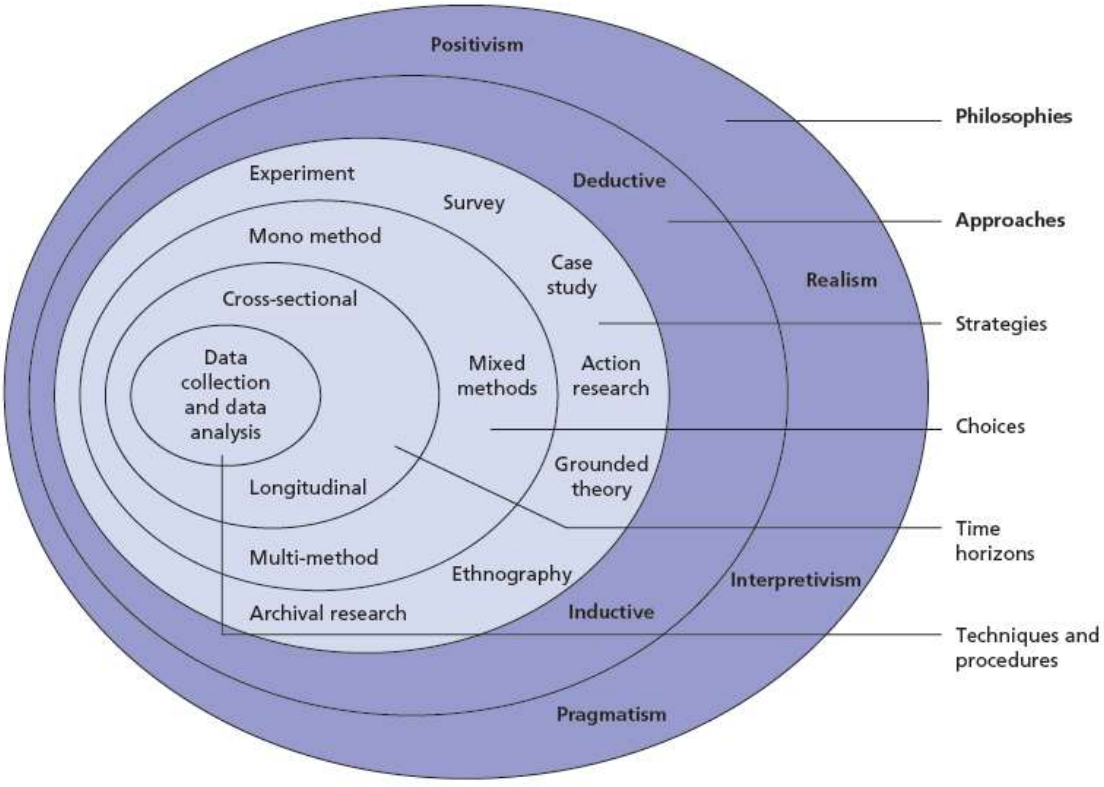


Figure 2-1: The research onion

a. Techniques and procedures

This study applied the data collection and data analysis techniques and procedures.

The data analysis will be done in Chapters 4 and 5 of the study.

b. Time horizons

There are two time horizons that can be applied to any research, namely the cross-sectional and the longitudinal.

a. Cross-sectional: Lewis-Beck et al. (2003) states that a cross-sectional design can use both qualitative and quantitative research, as they both measure an aspect or behaviour of many groups or individuals at a single point in time. A cross-sectional survey collects data to make inferences about a population of interest at one point in time.
b. Longitudinal: Similarly to cross-sectional design can also use quantitative and qualitative research, but the difference is that they study events and behaviours using concentrated samples over a long period (Lewis-Beck et al., 2003). Longitudinal research is used to find relationships between variables that are not related to a lot of background variables. It also involves studying the same group of individuals over an extended period, and also allows to study changes over time(Lewis-Beck *et al*., 2003).

Therefore this study makes use of longitudinal research with the aim of determining the pregnant woman's risk of HIV over ten years.

c. Choices
    a. Mono method research: This current study made use of mono methods, which is known as when either quantitative or qualitative data is collected rather than a combination of both (Saunders *et al.*, 2009). This study made use of quantitative on coded data.
d. Strategy
    a. Lewis-Beck et al. (2003) state that a survey is often associated with a deductive approach, and that it provides an economical way of collecting large amounts of data to address any given topic. This study made use of 10 annual survey samples of HIV antenatal data. Section 2.6 discusses data collection.

e. Approaches

There are two approaches that can be used, namely deductive and inductive approach.

a. Inductive approach: Saunders et al. (2009) refers to inductive research approach as the building theory. It allows for human aspects such as feelings and perceptions to be considered, other than facts. The collected data is used to understand a problem and to formulate a reasonable explanation (Lewis et al., 2007).

b. Deductive approach: Deductive reasoning argued that knowledge is gained by formulating a general statement and refining the statement by using logical arguments, which will then lead to a logical conclusion (Saunders *et al*., 2009). Deductive reasoning is applied where a theory is formulated and data are collected to either support or reject the theory, and is normally associated with positivism and realism (Lewis *et al.*, 2007).

Deductive research approach has the following characteristics (Saunders et al., 2009):

a. An urge to explain casual relationships between variables.
b. Quantitative data collection mostly takes place.
c. Control measures are put in place to allow the testing of hypotheses.
d. A structured methodology is followed.
e. The researcher is independent of what is being tested.
f. Large enough sample sizes are used to allow generalisation to be applied.

In this study, a deductive research approach was followed. Factors and relationships between factors were studied to determine their effect on the risk of HIV.

The sections below further explain the research philosophy and paradigm, and methodology used.

2.2 RESEARCH PHILOSOPHY

Research philosophy is the development and continuous improvement of knowledge as well as the nature of the knowledge (Saunders et al., 2009).

There are three well-known research philosophical aspects which Saunders et al. (2009) identify: epistemology, which describes what is acceptable knowledge in research; ontology, which is the study of the nature of knowledge; and axiology, which is the study of judgement about values.

A discussion on the research paradigms follows in order to position this study.

2.3 RESEARCH PARADIGMS

Scotland (2012) identifies three research paradigms, namely social constructionism, interpretivism and positivism.

A fourth research paradigm has been introduced which is known as the Design Science Research paradigm (DSR). Vaishnavi and Kuechler (2004) explain it as a paradigm that introduces the development of artefacts to solve problems.

According to Peffers et al. (2007), DSR is a diligent process of designing artefacts to solve identified challenges or problems, to contribute to research, to evaluate the designed artefacts and communicate their results to the relevant viewers.

Hevner et al. (2004) state that through the creation of new and innovative artefacts, DSR seeks to extent the boundaries of human organisational capabilities.

Constructionists or critical researchers state that social reality has always been present in the form of history, and is produced and reproduced by people (Aliyu et al., 2014). Wahyuni (2012) states that researchers who follow the social constructionist paradigm are part of the research, meaning that they cannot be separated from the truth and are therefore subjective.

The selection of the statistical method in the present study is restricted by the coded data set.

According to Aliyu et al. (2014), interpretivists posit that there are multiple methods of acquiring knowledge and that there is not just a single worldwide or universal truth. Research in this paradigm is conducted through the use of case studies, field experiments, exploratory analysis and qualitative analysis, and the research is directed at understanding the world or the truth from the individual’s perspective (Scotland, 2012).

Individual philosophies are explained and understood through interaction between researcher and participants (Guba and Lincoln, 1994), which means that interpretivists believe that knowledge and truth are discovered by interacting with the world and being conscious of one's surroundings (Scotland, 2012).

In the present study the risk of the mother having HIV was estimated from demographic variables, and the estimate depended on the model and the variables used.

Krauss (2005) states that positivists' core argument is that the social world exists externally from the researcher. Positivists are concerned with attempting to identify causes that affect outcome (Scotland, 2012), and they believe that knowledge is acquired through experience of the senses and can be attained through observations and experiments (Noor, 2008).

Positivism focuses on the gathering of quantitative data which is analysed with the use of statistical methods, with some focus on the relationship between the variables (Aliyu et al., 2014).

2.4 RESEARCH METHODS

One of the most important elements that goes into research is the specific method of data collection and analysis, which can be collected in various ways such as using an instrument or test, a behavioural checklist, or by visiting a research site and observing people's behaviours without talking or interviewing them about that particular subject (Creswell, 2013).

2.4.1 Three approaches to research

There are three main approaches to research, namely the quantitative, qualitative and mixed method approach. Creswell (2013) explains them as follows:

a. Quantitative approach: This is an approach in which the researcher uses positivist claims of acquiring knowledge through the use of cause and effect, measurements and observation. This approach makes use of experiments and surveys and predetermined instruments that assist in yielding statistical data.
b. Qualitative approach: This is an approach in which the inquirer makes knowledge claims based mainly on the constructionist view, such as the use of

ground theory studies and case studies. In this approach data is collected with the purpose of developing themes from the data.

c. Mixed method approach: In the mixed method approach knowledge is based on pragmatic grounds by collecting data either simultaneously or sequentially to better understand research problems. The data collected is both numerical information as well as text information, so that the final records represent both quantitative and qualitative information.

Figure 2.2 gives a summary of the research approaches and the various methods used.

Research Approach	*Knowledge Claims*	*Strategy of Inquiry*	*Methods*
Quantitative	Postpositivist assumptions	Experimental design	Measuring attitudes, rating behaviors
Qualitative	Constructivist assumptions	Ethnographic design	Field observations
Qualitative	Emancipatory assumptions	Narrative design	Open-ended interviewing
Mixed methods	Pragmatic assumptions	Mixed methods design	Closed-ended measures, open-ended observations

Figure 2-2: Research approach, knowledge claims, strategy of inquiry and methods (Creswell, 2013).

2.5 DESIGN OF EXPERIMENTS

This section gives a brief background of the origin of Design of Experiments (DOE) as well as its fundamental principles. It also discusses different uses of DOE and the components that make up DOE, such as the factorial design.

2.5.1 Brief background to Design of Experiments

DOE, also referred to as experimental design, is described by Telford (2007) as a structured and orderly manner of conducting an experiment as well as a method of analysing how the factors in question affect the outcome of the response.

DOE was invented by Ronald A. Fisher in the 1920s in his Rothamsted laboratory. He had initially invented DOE for agricultural use, but the procedure has found its way into the military and numerous scientific fields. It enables designers to determine concurrently the individual as well as the interactive effects that more than one factor could have on the output of a design (Telford, 2007).

Oehlert (2010) states that an experiment is identified by the treatments or factors as well as by the experimental units that are used. It is also recognised by the way the treatments are allocated to units as well as the responses that are measured.

In this study the factors are the pregnant woman's demographic variables.

2.5.2 Advantages of Design of Experiments

DOE offers certain advantages to experimenters. According to (Oehlert, 2010):

a. DOE allows the flexibility of comparing more than one treatment of interest.
b. DOE enables the design of experiments to minimise any form of bias in the treatments being compared.
c. Experiments can be designed to minimise errors in comparison.

DOE gives the experimenter control over experiments, which allows the experimenter to be able to make stronger inferences concerning the nature of variations in the experiment.

In this study the experimenter does not have control over a pregnant woman's demographic characteristics.

2.5.3 Main uses of Design of Experiments

There are numerous uses of design of experiments, but Telford (2007) states the following as the main uses:

a. *Discovering interactions among factors*

An interaction happens when the effect on the response of a change in the level of one factor depends on the level of another factor. When an interaction occurs between two factors, the combined effect of these particular factors on the response variable cannot be determined from the factors separately, and the effect of these combined factors can either be greater or lesser than that of the factors separately.

b. Screening many factors

Screening designs are used when there is a need to evaluate a process that has many factors with measured output variables. Using screening designs assists in determining which factors have the greatest effect on the response variable, for example, screening design in this study was used with the aim of determining which of the pregnant woman's demographic characteristics had an effect on the risk of HIV. Screening designs mostly consist of two-level factors and can also be referred to as characterisation testing or sensitivity analysis.

c. Establishing and maintaining quality control

A process is considered to be out of statistical control when either the mean or the variable is out of the specified controls. When this occurs the cause needs to be identified and rectified, and experimental design is very useful, similar to the screening design, except that there need not be two levels for all the factors.

d. Optimising a process

Optimising a process means determining the shape of the response variable. A screening design is normally used in this instance to determine which factors are most important. A response surface design has numerous levels on each of the factors, which assists in providing a clearer picture of the surface as well as providing information on which factors have curvature, and on which areas in the response peaks and plateaus occur.

e. Designing robust products

Designing robust products means learning how to cause the response variable to be unresponsive to uncontrollable inconsistencies in manufacturing processes.

2.5.4 Fundamental principles of DOE

Every design or technique consists of principles that are at the core or centre of what the technique describes or is made up of. The following section describes the fundamental principles that make up DOE (Telford, 2007):

a. Randomisation

Randomisation prevents unknown bias from distorting the results of the experiment, as well preventing one's personal and systematic biases from being included in the experiment (Gupta and Parsad, 2006).

In this study the dataset may be viewed as a random sample of the population each year.

b. Replication

Replication increases the initial sample size and is a technique that is useful for increasing accuracy within an experiment. Gupta and Parsad (2006) define replication as the repetition of the factors (treatments) under investigation to different experimental units, and is vital to ensure that the experiment is accurate.

c. Blocking

Blocking is a process of eliminating known nuisances so as to increase the accuracy of the experimental results.

d. Orthogonality

Orthogonality is described as an experiment resulting in the factor effects being uncorrelated and therefore being easier to interpret. The factors in an orthogonal experiment design are varied independently of each other.

In this study the factors were not varied but observed, meaning that the factors were not assumed to be independent.

There are numerous designs available in DOE, and although this study will only use two-level full factorial models, a brief description of the different designs was provided for literature purposes.

Numerous designs are available in DOE, namely:

a. *Response surface design:* This is a design that consists of lesser amounts of continuous factors, and is mainly used when the experimenter is certain about which factors are most important. Response surface design creates a predictive model of the relationship between the factors and the response (JMP, 2014).

b. *Split Plot design:* This is used when it is convenient to run an experiment in groups, and where one or more factors remain constant in each group (JMP, 2014).

c. *Screening designs:* These are the most popular designs and are mainly used when an experimenter wants to determine which factors in an experiment have the greatest effect on the result of the experiment, and require very few experimental runs. (JMP, 2014:101).

d. *Mixture designs:* According to JMP (2014), mixture designs are used for factors that are part of an ingredient in a mixture.

Although there are numerous designs available in DOE, for the present study factorial design was selected as the focus. The next section discusses factorial designs.

2.5.5 Factorial designs

Factorial experiments investigate the effects of two or more factors on the output. The present study investigated the effect that the pregnant woman's demographic factors had on the risk of HIV.

Factorial experimentation is a method in which factors as well as the combination of factors are measured (Telford, 2007, Mee, 2009)

Within factorial design is the full factorial design which considers all possible combinations of the factor levels (JMP, 2014). The full factorial design is considered to be very accurate due to the fact that it performs an experimental run at every combination of the factor run, and is therefore more time consuming and costly (Bingöl et al., 2015). A fractional factorial design only looks at a subset of the experimental runs of a full factorial design (Bingöl et al., 2015).

A two-level full factorial design is denoted as 2 to the power k, where 2 is the number of levels and k is the number of factors in the experiment (Anderson and Whitcomb,

2015). For example, if we have K factors each at two levels, the full factorial consists of 2x2x...x= 2^K combinations (Mee, 2009). The pregnant woman has six demographic characteristics which are the factors considered in this study, and each factor has two levels and therefore the full factorial consists of 2X2X2x2x2x2 =2^6 combinations.

Two-level designs are well known and are used in many applications, particularly when there are many factors to be considered. They are also primarily used in studies where the main purpose is to determine which factors have the greatest influence on the response variable, and not necessarily which combination might be most optimal (Morris, 2011). The study also seeks to determine which of the pregnant woman's demographic characteristics influences the risk of HIV.

Mee (2009) states that some of the benefits of using factorial designs is that they reveal whether the effect of each factor depends on the level of another factor, and helps formulate linear models which summarise the combined effect of the factors well.

Within a two-level full factorial model, aside from the main effects, factors can result in interaction effects, which are caused by two or more factors interacting with each other, and these can cause main effects to be insignificant. Therefore factorial experiments can be defined as experiments in which both the main effects and interactions of more than one factor are studied together (Morris, 2011).

Factorial models allow the study of individual effects of each factor, as well as the effect of the interactions, using less resources and money (JMP, 2014).

Cavazzuti (2013) states that the main and the interaction effect give a valuation of the effect the factors , or the interaction of the factors has on the response variable.

An advantage of a full factorial model is that it uses the data very efficiently and does not confound the effects of the parameters, therefor making it easier to evaluate and analyse the main and the interaction effects clearly (Cavazzuti, 2013).

2.5.1.1 Two-level model

The pregnant woman's demographic characteristics were split into two levels as presented in Table 2.1 primarily because there were two parts to the demographic characteristics being studied. The format given below of the two levels of the pregnant woman's demographic characteristics were applied throughout the study.

Table 2-1: Factors and levels table

	Levels	
Factors	**-1**	**1**
Mother's age	<= 24	>24
Father's age	<= 28	>28
Education (grades)	Primary	Secondary and tertiary
Gravidity (number of pregnancies)	1	>1
Parity (number of children)	0	>1
Syphilis	0	1

The demographic characteristics were defined in chapter one.

2.5.6 Components of DOE

The components of an experiment or DOE include treatments, experimental units, responses as well as a method used to assign units to treatments. The section below briefly explains the components of DOE as well as the terms used in DOE (Oehlert, 2010: 6 - 8).

a. Treatments are defined as the different components that will be compared in an experiment.

b. Experimental units are classified as those that are applied to the treatments.

c. Responses or a response variable are the outcome of the effect of the treatment, for example the response variable in this study is HIVrisk, and may changes per the effect of the factor.

d. Experimental error is the random variations found in all experimental designs.

e. Measurement units or response units are defined as the objects on which the response is measured. In the present study pregnant women were studied.

f. Blinding occurs when the evaluators of the response do not know to which treatments which units allocated. Blinding assists in preventing bias.

g. Confounding or a confounding rule is declared when the effect of one factor cannot be separated from that of another factor, except in a special condition where confounding should be avoided.

h. An effect is defined as a change in the response variable resulting from changes in the factor level.

In present study if mother's age, education level, gravidity, syphilis or any of the other factors changes, it may affect the response which is the risk of HIV. A change can either cause a positive or a negative effect to the response variable, which means an increase or decrease in the risk of HIV.

2.5.7 Analysis of variance

Analysis of variance also known as Anova is a multivariate method used to analyse variation in a response variable normally used to test equality among means by comparing variance among groups relative to variance within groups (Larson, 2008).

Anova was perfected by Ronald Fisher by using it to analyse results of agricultural experiments, but today Anova is widely used in the field of research (Larson, 2008).

Analysis of variance uses the following quantities, each used to measure various kinds of variation in test statistic (Swanepoel et al., 2011).

Analysis of variance makes it possible to summarise data so that relationships and patterns can be easily interpreted and understood (Yong and Pearce, 2013).

Moore et al. (2012) state that the advantages of anova are as follow:

a. Valuable resources can be spent more efficiently by studying two factors simultaneously rather than separately.
b. The residual variation in a model can be reduced by including a second factor thought to influence the response
c. Interactions between factors can be investigated.

The definition of interaction is that the effect of a change in the level of one factor on the mean outcome depends on the level or value of the other factor, therefore an interaction term is part of a statistical model (Seltman, 2012).

Analysis of variance is further explained in chapter 3.

The next section discusses data collection.

2.6 DATA COLLECTION

Data collection methods or techniques yield data about people, objects, phenomena and the environment in which they occur to be collected in a systematic way (Chaleunvong, 2009).

There are various data collection techniques, namely (Chaleunvong, 2009, Saunders et al., 2009):

a. *Using available information* allows the use of information that has already been collected by someone else; the information might not yet have been published or analysed.
b. *Observing* involves systematically selecting, watching and recording the behaviour or characteristic of a person or an object.
c. *Interview* involves asking questions and receiving response from an individual or a group.
d. *Questionnaires* are a data collection technique in which questions are presented to the respondents to answer in written form.
e. *Focus group* is a technique in which a group of 8-12 people have a discussion about a particular subject under the guidance of a facilitator or reporter.

As mentioned in Chapter 1, the data used in this study was collected by the NDoH, which conducts annual antenatal HIV prevalence unlinked surveys targeting pregnant women attending antenatal clinics in the public health sector (NDoH, 2012).

The NDoH uses of two selection criteria, namely inclusion criteria and exclusion criteria:

a. *Inclusion criteria*: Are the characteristics the subjects should have to be included in the study. In this case it describes all pregnant women attending antenatal clinics for the first time during their current pregnancy.

b. *Exclusion criteria*: Are the characteristics which disqualify the subject from the study. In this case it describes pregnant women that had previously attended an antenatal clinic during their current pregnancy.

The two selection criteria were used to avoid duplication within the data.

The sample collection described by the NDoH (2012) is that a full blood analysis was carried out on pregnant women attending antenatal care for the first time during their current pregnancy as an entry point for HIV testing using anonymous unlinked procedures. The blood was labelled with a bar code. The pregnant woman's demographic characteristics are collected in such a way that it is not possible to ascertain the identity of the patient using a standardised data collection form. This information is then marked with the same bar code used for the blood sample.

Therefore the present study used available data. Coded data were used as this was the only data made available to the researcher.

2.6.1 CHAPTER CONCLUSION

The objective of this chapter was to gain an understanding of the research methodology, and focused on the design of experiments.

The objective of investigating the research philosophy, research paradigm and research approaches was achieved.

DOE methodology was used in this study because the objective was to develop a two-level full factorial model, which takes into consideration all the factors and not just the subset of the factors.

The two-level full factorial design was chosen as the pregnant women's demographic characteristics had two levels each.

The chapter also discussed the various components of DOE, and gave definitions of a factor, an experimental unit and a response variable.

Chapter 2 also discussed the different data collection techniques, focusing on the technique used by the NDoH to collect HIV data on pregnant women attending antenatal clinics. The chapter also gave a definition of the different demographic

characteristics of the pregnant women and the process that was used in the research.

Chapter 3 briefly describes literature on statistical methods with the aim of gaining a better understanding of the statistical methods related to this study.

CHAPTER 3: STATISTICAL METHODS

3.1 INTRODUCTION

In chapter 2 the study objective was to demonstrate an understanding of research methodology and how it contributes to the development of this study. The primary objective of this study was to develop two-level full factorial models to analyse antenatal HIV data. To achieve this, a search of the literature on statistical methods used in this study was required.

Isotalo (2001) describes statistics as a method that is used to collect, analyse, interpret and formulate conclusions from information provided or collected.

In this study, statistics was used to analyse antenatal HIV data to better understand the risk of HIV of a pregnant woman.

Peck *et al.* (2015) defines statistics as a science that puts close attention on collecting, analyse and drawing conclusions from data.

The objective chapter is to demonstrate an understanding of statistical methods, and how it contributes to the development of this study.

The chapter is divided into the following sections: History of the data (Section 3.2), Contingency tables (Section 3.3), Regression analysis (Section 3.4), Simple linear regression (Section 3.5), Multiple linear regressions (Section 3.6) and the conclusion (Section 3.7).

3.2 History of the data

This section examines the history of the data and HIV studies conducted in countries such as Tanzania and Ethiopia, and the trends that have been found to be prevalent in those countries. As stated previously, the Sub-Saharan region has the most HIV cases in the world, therefore other countries on the African continent took it upon themselves also to conduct surveys to assist them to monitor the HIV epidemic and find ways to combat it.

Research conducted by UNAIDS revealed that Sub-Saharan Africa is the region with the highest incidence of HIV/AIDS infection (NDoH, 2014). In the light of this, the National Department of Health (NDoH) introduced a new way of monitoring the disease by introducing a yearly nation-wide HIV survey.

The yearly national prenatal HIV prevalence survey is conducted among pregnant women attending their first appointment at a public clinic. The survey is conducted in October in all nine provinces in 52 health districts. A cross-sectional standard unlinked and anonymous survey is conducted among pregnant women of ages 15 to 49. The survey has assisted the NDoH to monitor HIV and syphilis prevalence trends since 1997 (NDoH, 2012).

As mentioned before this study makes use of coded antenatal data of pregnant women, because this was the only data available to the researcher.

The demographic characteristics of the pregnant woman which are the variables of interest are described in CHAPTER 2 under TABLE 2.1.

A study of the prevalence of syphilis and HIV was conducted among pregnant women who attended the University of Gondar teaching hospital in north-west Ethiopia. The aim of the study was to determine the effect of syphilis on acquiring HIV (Endris et al., 2015).

According to Endris et al. (2015), a cross-sectional study was conducted for the period from February to June. Of the 385 pregnant women who took part in the study, 11 tested positive for reactive syphilis, 43 tested positive for HIV and 2 tested positive for both HIV and syphilis. Owing to these findings, the study concluded that HIV and syphilis infections were still prevalent in Ethiopia and that re-screening was necessary for all pregnant women during antenatal care.

According to Lawi *et al*. (2015), pregnant women in Tanzania are only tested during their antenatal care, and this has resulted in missed opportunities of re-screening for HIV and syphilis of women after giving birth. Therefore a cross-sectional hospital-based study was conducted among pregnant women attending antenatal care at the Bugando Medical Centre from January to March 2012.

The study revealed that of 331 pregnant women who had tested negative for syphilis during their antenatal care screening, 9 (2.7 %) tested positive for syphilis at delivery, and of 331 pregnant women who had tested negative for both syphilis and HIV during antenatal screening, 8 (2%) tested positive at birth. Therefore the study concluded that re-screening at birth is important so as not to overlook women who might have contracted syphilis and HIV during pregnancy (Lawi et al., 2015).

As stated in the problem statement, the gap in literature that the present study intends to fill is to develop two-level full factorial models with which to analyse antenatal HIV data. This study took into consideration all the demographic factors of the pregnant women and analysed their risk of acquiring HIV.

3.3 Contingency tables

Understanding and describing the data you have is one is important in a statistics (Lawal, 2014), therefore the next steps after collecting data is organising it so that it is easy to read and understand, as well as see trends if any exists (Manikandan, 2011).

One of the widely used methods is frequency distribution. Frequency distribution is defined as an organised table of the number of individuals located in each category

on the scale of measurement(Swanepoel et al., 2011).It allows researcher to have a better view of the data, and presents a picture of how the individual observations are distributed in the measurement scale.

Frequency distribution is mostly discussed for quantitative or qualitative single variable data set, from which the data is summarised and presented in a frequency table (Steyn and Swanepoel, 2008). However the data of interest in this study is categorical in nature. For example pregnant women of two age groups, young and older, it is of interest to know: Are they HIV negative positive?

Categorical data can be cross-clarified to get a count of the number of cases with the same combination of levels, by creating a multi-way contingency table showing the levels and the counts.

This study used contingency tables to better understand the demographical characteristics of the pregnant women, and used the results in chapter 4 for trend analysis.

Steyn and Swanepoel (2008) describe a contingency table as a table that lists the number of counts for a joint occurrence of two or more levels or possible outcomes, one level for each of the categorical variables

A 2x2 pronounced 2 by 2 table was used because the demographical characteristics of the pregnant woman are two categorical variables each with two categories. In a cross-tabulation, one variable will be the row variable and the other will be the column variable (Stokes et al., 2012).

Table 3-1: Contingency table of treatment.

	No disease	Type 1 disease	Type 2 disease	Totals
Treatment	200688	24	33	200 745
Placebo	201,087	27	115	201 229
Totals	401,775	51	148	401 974

TABLE 3.1 is an example of a contingency table of two categorical variables that was used for a trial of treatment status, which had two levels; treatment and placebo, and

the disease status, which had three levels, no disease, type 1 disease and lastly type 2 disease.

TABLE 3.1 shows that of 200,745 individuals who were treated 24 contracted type 1 disease and 33 contracted type 2 disease, and 200,688 did not contract any disease.

The above results show that of 201,229 of the individuals who received the placebo, 27 contracted type 1 disease and 115 contracted type 2 disease, and 401,775 did not contract any disease.

As mentioned above contingency tables were used in this study to better understand the demographic characteristics of the pregnant women.

Contingency tables were formulated for each of the six variables for the 10 year period, refer to APPENDIX A.

TABLE 3.2 shows the results of the contingency table of HIVstatus against the pregnant woman's age for the year 2001.

In the table below the variable HIV class has two disjoint categories namely HIV negative (0) and HIV positive (1). The variable Mother's age (Mothage) also has two disjoint categories namely young pregnant women(-1) and older pregnant women(1).

Table 3-2: Contingency table of HIV class by mother's age 2001

HIVclass by Mothage 2001			
HIVstatus	**Mothage**		
	-1	**1**	**Total**
0	4506 77.00%	5002 78.18%	9508
1	1346	1396	2742

HIVclass by Mothage 2001			
HIVstatus	**Mothage**		
	-1	**1**	**Total**
	23.00%	21.82%	
Total	5852 47.77%	6398 52.23%	12250 100.00%

The inside of the table is called the joint distribution of the two variables, and the lower row total of 5852, 6398 together with the total 12250 make up the marginal distribution of pregnant woman's age. Similarly the column total of 9508, 2742 together with the total 12250 is the marginal distribution of the pregnant woman's HIV status. The table shows that for the year 2001 the risk of young pregnant women of contracting HIV was 23% and the risk of older pregnant women contracting HIV was 21.82%.

The results from the contingency tables in APPENDIX A were used in chapter 4 to determine the effect of the pregnant woman's demographic characteristics on the risk of HIV.

In this study the contingency tables were calculated (see appendix A) and will be used in CHAPTER 4.

3.4 REGRESSION ANALYSIS

Regression analysis enables questions concerning data to be answered, and for patterns in the data to be discovered. Allen (2007) states that regression analysis can be dated back to the late nineteenth-century England to a scientist named Francis Galton. This important insight was discovered when Galton was studying how the human characteristic of height was passed on from one generation to the next. He did this by collecting samples on the height of individuals and the height of their parents. Galton's study concluded that tall people usually had tall parents, although they would not be as tall as their parents, and shot people usually had short parents although they

would not be as short as their parents. Regression analysis techniques has since then been used by researchers to study various types of data.

Regression analysis is a quantitative research method which is used to model and analyse variables where the relationship has a dependent variable and one or more independent variables(Campbell and Campbell, 2008). One of the many reasons regression analysis is used is to study and understand the relationship between independent variable and the dependent variable (Montgomery, 2017). This study seeks to study and understand the effect that the pregnant woman's demographic characteristics have on her risk of HIV.

Rawlings et al. (2001) defines modeling as the development of mathematical expression that describes the behaviour of a variable or variables of interest. The variables can range from the price of petrol, or in the case of this study the risk of HIV. The variables are are called dependent variables and denoted with (Y), and the modeling is most commonly aimed at describing how the mean of the dependent variable changes.

The independent variables denoted by X, are described to be explanatory or predictor variables with subscripts needed to identify different independent variables (Rawlings et al., 2001).

Regression can be used to show the relationship between one independent variable and a dependent variable, as formulated below:

$$Y = Bo + B1x + u$$

Campbell and Campbell (2008) state that the magnitude and direction of the relation is given by the slope parameter denoted (B1), and the status of the dependent variable when the independent variable is absent is given by the intercept parameter (Bo). The error term (u) determines the amount of variation not predicted by the slope and the intercept term, and the regression coefficient R-square shows how well the values fit the data.

2.5.2 Main uses and advantages of regression analysis

As mentioned above, regression analysis is used to estimate the relationship between two or more variables, and also offers a number of benefits namely (Ray, 2017):

f. Discovering interactions among factors

An interaction happens when the effect on the response of a change in the level of one factor depends on the level of another factor. When an interaction occurs between two factors, the combined effect of these particular factors on the response variable cannot be determined from the factors separately, and the effect of these combined factors can either be greater or lesser than that of the factors separately.

g. Identifying relationships

Regression analysis is used to indicate significant relationships between dependent variables and independent variables.

h. Strength of variables

Regression indicates the strength of impact of multiple independent variables on a dependent variable.

Regression analysis also allows for the comparison of the effect of variables measured on different scales. These benefits assist market researchers, data analysts and data scientists to eliminate and evaluate the best set of variables to be used for building predictive models (Ray, 2017).

2.5.4 Applications of regression analysis

There are three main uses of regression analysis, namely causality, forecasting and prediction (Gogtay et al., 2017)

e. Causality

Causation is known to indicate relationships between two events, where one is affected or has an effect on the other (Rawlings et al., 2001). This study seeks to determine how the change in the pregnant woman's demographical characteristics affects the risk of HIV.

Gogtay et al. (2017) uses an example of a study that was conducted on working aged people from the general population in the United Kingdom, to estimate the risk of occupational exposure to noise on self-reported hearing difficulties using a validated questionnaire.

The study found that in both male and female the risk of severe hearing difficulty increased with years spent working in a noisy job.

f. Forecasting

Gor (2009) defines forecasting as a process of making predictions about the future based on past and present data, and by analysing the trends that emerge from the data.

Gogtay et al. (2017) gives an example from a study on efficient management of patient process in the emergency department in a hospital, by studying diverse models in an attempt to forecast the daily number of patients seeking emergency department services using calendar variables and ambient temperature reading as the independent variable.

The study found that the mean number of emergency department visits was 389 with a seasonal distribution, with the highest patient volume seen on Monday and lowest on weekends.

This study seeks to determine whether there is a trend by taking a closer look at the pregnant woman's demographical factors over time in chapter 4.

g. Predictions

Gogtay et al. (2017) used an example of a study that was conducted to predict risk factors for colorectal cancer in a community practice where they studied 461 consecutive patients undergoing colonoscopy. 293 patients were randomly selected and they evaluated the impact of several independent variables in a model that looked at prediction of occurrence of colorectal cancer.

The five variables used in the study were, the patient's age, gender, haematocrit, fecal occulant blood test results and indication for colonoscopy. When the model was applied to the remainder of the 169 patients it was found to be a reliable indicator of the risk of colorectal neoplasia.

2.5.5 Types of regression

There are numerous types of regression, namely simple linear regression, multiple linear regression, logistic regression and polynomial regression, but for the purpose of this study we will only look at simple linear regression and multiple linear regressions.

e. *Simple Linear regression:*

Gogtay et al. (2017) defines simple linear regression as the most commonly used regression technique. Simple linear regression is used when there is a single dependent and single independent variable, where both the variables must be continuous and the line describing the relationship is called a straight line.

f. *Multiple linear regression:*

Multiple linear regression is used when there is one continuous dependent variable and two or more independent. The variables can be quantitative or qualitative, and can be presented either as continuous data or qualitative data (Gogtay et al., 2017).

3.5 SIMPLE LINEAR REGRESSION

Simple linear regression assists in studying the relationship between a response variable denoted Y, and an explanatory variable denoted x (Moore *et al*., 2012).

The relationship determines the amount of change in one variable that is associated with the change in another variable or variables(Gogtay *et al.*, 2017).

In this study the response variable is HIV risk and the explanatory variables are the pregnant woman's demographic characteristics.

The simple linear regression model is as follow (Moore *et al*., 2012):

Given *n* observations on the explanatory variable *x* and the response variable *y,*

$$(x1, y1), (x2, y2), \ldots, (Xn, Yn)$$

The statistical model for simple linear regression states that for each *i* from 1 to *n* the observed response is as follow:

$$Y1 = Bo + B1Xi + ei$$

Where $Bo + B1Xi$ is the mean response when $x = xi$. The deviations Ei are assumed to be independent and normally distributed with mean 0 and standard deviation.

Rawlings *et al*. (2001) states that the method of least-square explains the relationship between the explanatory variables and the response variable. The section below gives a brief overview of method of least-square.

3.5.1 Method of least-square

The least-square line is obtained by minimising the sum of squares of the vertical distances between the observed points and the corresponding points on the line (Gogtay et al., 2017).

Suppose a straight line is to be fitted through data points. The intercept and the gradient of the least-squares straight line $\hat{y} = a + bx$ are the values of a and b responsible for minimising the following expression:

$$\sum_{i=1}^{n}(yi - \hat{y}i)^2 = \sum_{i=1}^{n}(yi - (a + bx))^2$$

The least-square values of *a* and *b* can be proven to be:

$$b = \frac{\sum_{i=1}^{n} XiYi - \frac{1}{n}\sum_{i=1}^{n} Xi \sum_{i=1}^{n} Yi}{\sum_{i=1}^{n} Xi^2 - \frac{1}{n}(\sum_{i=1}^{n} Xi)^2}$$

$$a = \bar{y} - b\dot{x}$$

The straight line equation can be used to make predictions and forecasts. The linear straight line is used in chapter 4 to determine the effect that the pregnant woman's demographic factors have on the risk of HIV.

A regression model can also be used to forecast through interpolation and extrapolation.

Interpolation estimates a *y*-value for a give *x*-value inside the interval of observed x-value, and extrapolation estimates a y-value for a given x-value outside the interval of observed x-values (Swanepoel et al., 2011).

3.5.2 The coefficient of determination

The coefficient of determination determines how well the model fits the observed data (Swanepoel et al., 2011)

The measure of fit is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{\sum_{i=1}^{n}(yi-\bar{y})^2}$$

The coefficient of determination has the following properties:

a. $0 \leq R^2 \leq 1$.
b. R^2 =1 implies a perfect fit of the model to the observed data.
c. R^2 = 0 implies that the model does not fit the data.
d. For a straight line it is true that $r^2 = R^2$.

3.5.3 Analysis of variance for regression

Analysis of variance is used to summarise the information about the source of variation in the data, and is based on the Data = FIT + RESIDUAL framework (Moore et al., 2012).

The total variation in the response variable for example HIV risk, is described by the difference between $yi - \bar{y}$ (Moore et al., 2012, Rawlings et al., 2001). If the deviations are equal to 0, then all the observations are equal and there is no variation in the response (Gogtay et al., 2017).

According to Moore *et al*. (2012) the sum of squares added is as follow:

$$\sum_{(yi - \overline{y})^2} = \sum_{(\hat{y}i - \overline{y})^2} + \sum_{(yi - \hat{y}i)^2}.$$

With the equation written as

$$SST = SSM + SSE$$

Where

$$SST = \sum_{(yi - \overline{y})^2}$$

$$SSM = \sum_{(\hat{y}i - \overline{y})^2} \; and$$

$$SSE = \sum (yi - \hat{y}_i)^2.$$

The SS in the model abbreviation stands for sum of squares, and the T stands for total, the M stands for model and lastly the E represents the error.

The mean error sum of squares denoted MSE is defined as:

$$S^2 = \frac{\sum (yi - \hat{y}_i)^2.}{n-2}$$

The mean square MS is denoted as follow

$$MS = \frac{sum\ of\ squares.}{degrees\ of\ freedom}$$

ANOVA calculations are displayed in an analysis of variance table, which has the following format for simple linear regression (Moore *et al.*, 2012)

Table 3-3: ANOVA table for simple linear regression

Source	Degrees of freedom	Sum of squares	Mean square	F
Model	1	$\sum (\hat{y}_i - \overline{y})^2$	SSM/DFM	MSM/MSE
Error	$n-2$	$\sum (yi - \hat{y}_i)^2$	SSE/DFE	
Total	$n-1$	$\sum (yi - \overline{y})^2$	SST/DFT	

The F column represents the test statistic for comparing the null hypothesis against the alternative hypothesis (Larson, 2008).

The p-value signifies the probability of a random variable having $F(1, n-2)$ distribution being greater than or equal to the calculates value of the F statistic (Moore *et al*., 2012).

3.6 MULTIPLE LINEAR REGRESSION

The section above explained how to analyse a linear relationship between a response variable and a factor. This current section will give a brief discussion of what multiple linear regressions are and how it will be used in this study.

The difference between simple linear regression and multiple linear regression is that multiple linear regression uses more than one factor to explain or predict a single response variable (Moore et al., 2012). In multiple linear regression the response variable y depends on not one but p explanatory variables or factors denoted $x1, x2, \ldots, xp$ *(Rawlings et al., 2001).* In this study the response variable is HIV risk and in chapter 5, we look at which of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

The statistical model for multiple linear regressions is *(Gogtay et al., 2017, Moore et al., 2012*)

$$Y = Bo + B1Xi1 + B2Xi2 + \ldots + BpXip + e$$

$$for\ i = 1, 2, \ldots, n.$$

the following assumptions hold (Larson, 2008, Seltman, 2012):

a. Independence

The value of one observation should not influence or affect the value of another observation.

b. Normality

The observed data was collected from a normally distributed population.

c. Homogenous variation

The population variation of the data within each group must be the same.

3.6.1 Analysis of variance for multiple regression

Similarly to simple linear regression, multiple linear regressions make use of an ANOVA table presented below (Moore *et al*., 2012).

Table 3-4: ANOVA table for multiple linear regression.

Source	Degrees of freedom	Sum of squares	Mean square	F
Model	p	$\sum(\hat{y}_i - \overline{y})^2$	SSM/DFM	MSM/MSE

Error	$n-p-1$	$\sum_{(yi-\hat{y}_i)^2}$	SSE/DFE	
Total	$n-1$	$\sum_{(yi-\overline{y})^2}$	SST/DFT	

The degrees of freedom of the model in the table above increases from 1 to p reflecting that more than one variable (Moore *et al*., 2012).

The sum of square is the source of variation, and the estimate of the variance is represented by the MSE in the ANOVA table (Larson, 2008).

Moore *et al.* (2012) state that the F statistic represented by MSM/MSE is used to test the null hypothesis,

$$Ho: B1 = B2 = \ldots Bp = 0$$

Against the alternative hypothesis

$$Ha = Bj \neq 0 \; for \; at \; least \; one \; j = 1,2,\ldots,p.$$

According to Gogtay *et al.* (2017) the null hypothesis denoted Ho says that none of the variables have an influence on the response variable when used in the form expressed by the multiple regression equation.

The alternative hypothesis denoted Ha states that at least one of the variables is linearly related to the response variable (Gogtay *et al*., 2017).

A large value of the F statistic gives evidence against the Ho, but if Ho is true the F value has the F(p,n-p-1) distribution (Moore *et al*., 2012).

In CHAPTER 5 two-level full factorial models are formed with the assistance of multiple linear regressions to determine which of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

3.7 CONCLUSION

The objective of this chapter was to provide an overview of the statistical methods that was applied in this study, and this objective was achieved by giving literature on the history of the data and a brief overview of similar studies conducted. This chapter also provided literature on contingency tables which was applied in CHAPTER 4 for the

analysis of the pregnant woman's demographical characteristics. This chapter gave a brief review on time series and linear models and lastly gave literature on multivariate analysis with a focus on Anova.

The following chapter is an analysis of the pregnant woman's demographical characteristics and their effect on the risk of HIV.

CHAPTER 4: DATA ANALYSIS

4.1 INTRODUCTION

The primary objective of this study was to develop two-level full factorial models for the analysis of HIV data. To achieve this, the demographic characteristics of pregnant women were investigated in order to understand the factors better.

The South African National Department of Health conducts an annual survey of the risk of a pregnant woman becoming infected with HIV. This is done by collecting their demographic characteristics, namely the pregnant woman's age, the father's age, gravidity, parity, level of education and syphilis status. This chapter seeks to better understand the demographic characteristics of the pregnant woman over the ten-year period and the changes that have occurred over time.

This chapter is divided into the following segments: data analysis (Section 4.2), differential effects (Section 4.3), and the conclusion (Section 4.4).

4.2 DATA ANALYSIS

The objective of this section is to analyse the pregnant women's demographic factors to understand the story behind each demographic characteristic of the pregnant women, and to determine whether the changes remain the same over time. The data used were restricted to coded data only, as this was the only data made available for the purpose of this study.

The year 2003 was found to have very little data, thereby causing it to be an outlier among the other data sets.

4.2.1 HIV risk to pregnant women on age

This section looks at mothers' age individually over the ten-year period. The analysis was done on the risk of HIV among pregnant women ages 24 years and younger, as well as the risk of HIV among pregnant women of ages 25 years and older. A conclusion is then given.

TABLE 4.1 shows the HIV risk of pregnant women of ages 13 to 24 from 2001 to 2010.

Table 4-1: HIV risk to young pregnant women from 2001 to 2010

Year	Coded Year	Total number of young pregnant women (ages 13 to 24)	Total number of HIV-positive young pregnant women	HIV risk Percentage
2001	1	5 852	1 346	23.00%
2002	2	7 692	1 804	23.45%
2003	3	924	249	26.95%
2004	4	7 445	1 869	25.10%
2005	5	6 695	1 634	24.41%
2006	6	16 090	3 582	22.26%
2007	7	16 615	3 668	22.08%
2008	8	16 692	3 611	21.63%
2009	9	15 797	3 410	21.59%
2010	10	15 221	3 279	21.54%

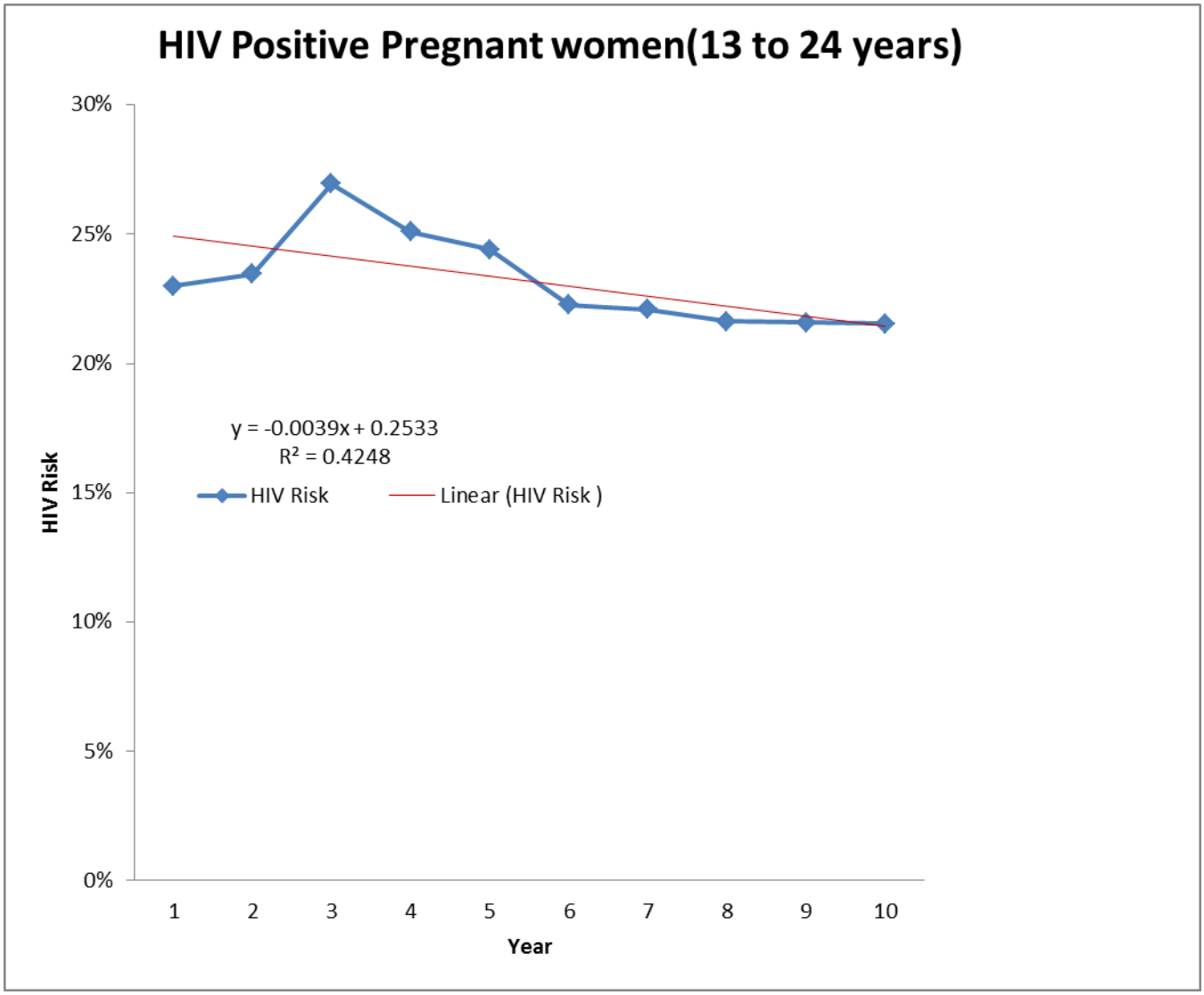


Figure 4-1: HIV risk to pregnant women of ages 13 to 24 years

FIGURE 4.1 shows a line graph of the risk of HIV for pregnant women of ages 13 to 24 years from 2001 to 2010. The graph shows a downward trend.

The linear model is as follows:

$$Y = -0.0039\,x + 0.2533$$

The negative slope shows a negative trend of the HIV risk of pregnant women aged 13 to 24 years over time. The risk of HIV decreased by -0.0039 yearly. The R-square shows that the model accounts for 42% of the variability of the response data around the mean.

Table 4-2: HIV risk to older pregnant women from 2001 to 2010

Year	Coded Year	Total number of older pregnant women aged 25 to 49	Total number of HIV-positive older pregnant women	HIV risk percentage
2001	1	6 398	1 396	21.82%
2002	2	7 641	2 273	29.75%
2003	3	946	255	26.96%
2004	4	7 371	2 493	33.82%
2005	5	6 095	2 119	34.77%
2006	6	15 168	5 407	35.65%
2007	7	15 866	5 754	36.27%
2008	8	15 879	5 859	36.90%
2009	9	15 346	5 724	37.30%
2010	10	14 891	5 744	38.57%

TABLE 4.2 shows the results of the HIV risk among pregnant women of ages 25 to 49 years from 2001 to 2010.

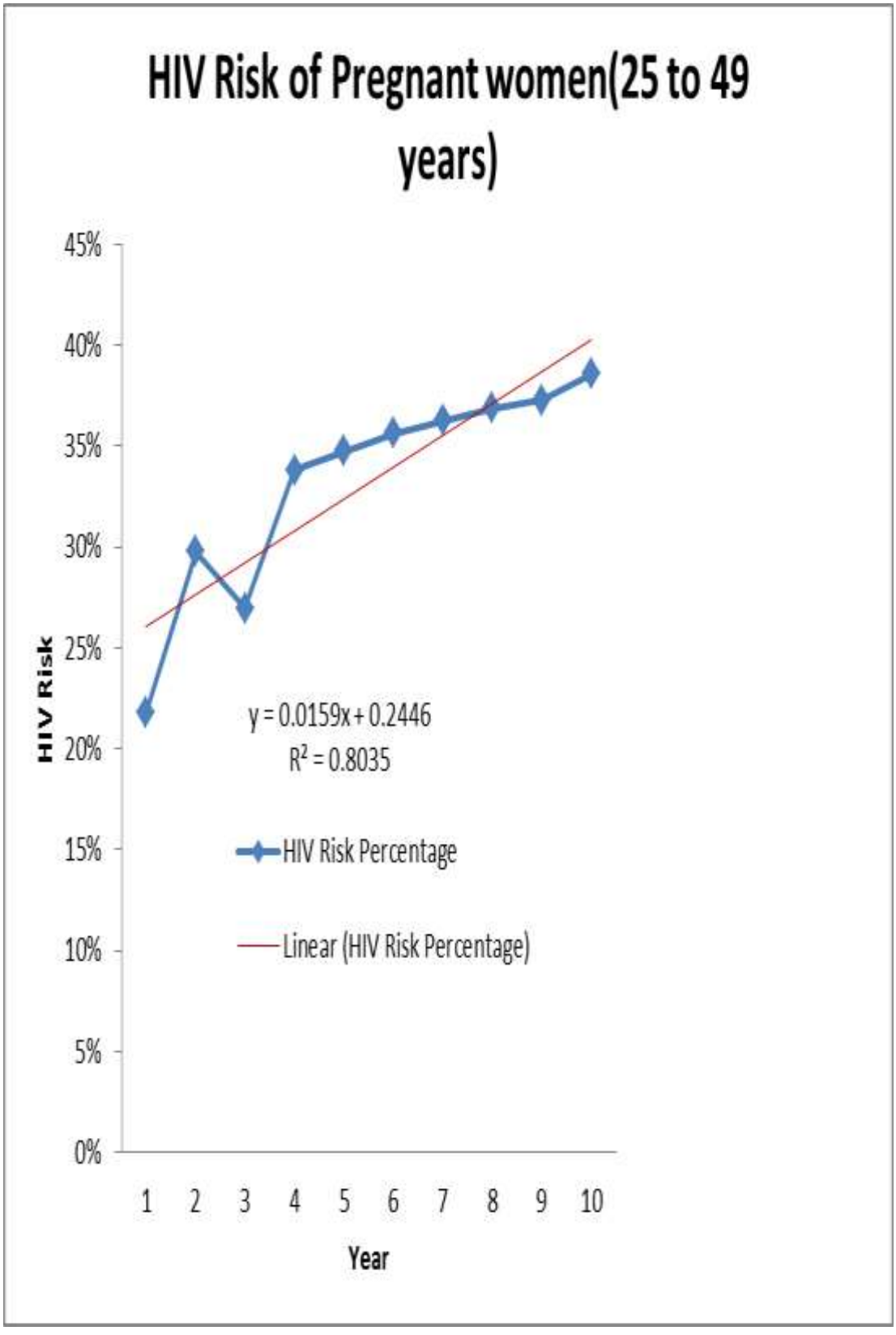


Figure 4-2: HIV risk to pregnant women aged 25 to 49 years

FIGURE 4.2 shows a line graph of the HIV risk of pregnant women aged 25 to 49 years. The graph shows an increasing or upward trend of the risk of HIV for older mothers. The linear model is as follows:

$$Y = 0.0159\,x + 0.2446$$

The positive slope shows an increasing trend of the risk of HIV of pregnant women aged 25 to 49 years over the ten-year period. The risk increases by 0.0159 yearly.

The model has an r-square of 0.8035, which signifies that the model explains 80% of the variability of the response data around the mean.

4.2.2 HIV risk to pregnant women whose partners are known

This section examines the HIV risk of pregnant women with partners aged 28 years and younger and partners aged 29 years and older. This section includes a time series graph that maps out the risk of HIV over the ten-year period with the aim of determining whether there is a trend.

TABLE 4.3 shows the HIV risk of pregnant women with partners younger than 28 years from 2001 to 2010. The table shows that the risk percentage slightly decreases yearly.

Table 4-3: HIV risk to pregnant women with partners 28 years and younger, 2001 to 2010.

Year	Coded Year	Total young partners	Total HIV-positive young partners	HIV risk percentage
2001	1	2 962	527	17.79%
2002	2	7 176	1 628	22.69%
2003	3	834	209	25.06%
2004	4	6 965	1 692	24.29%
2005	5	6 318	1450	22.95%
2006	6	15 089	3 292	21.82%
2007	7	15 865	3 428	21.61%
2008	8	16 160	3 464	21.44%

2009	9	15 140	3 226	21.31%
2010	10	15 098	3 259	21.59%

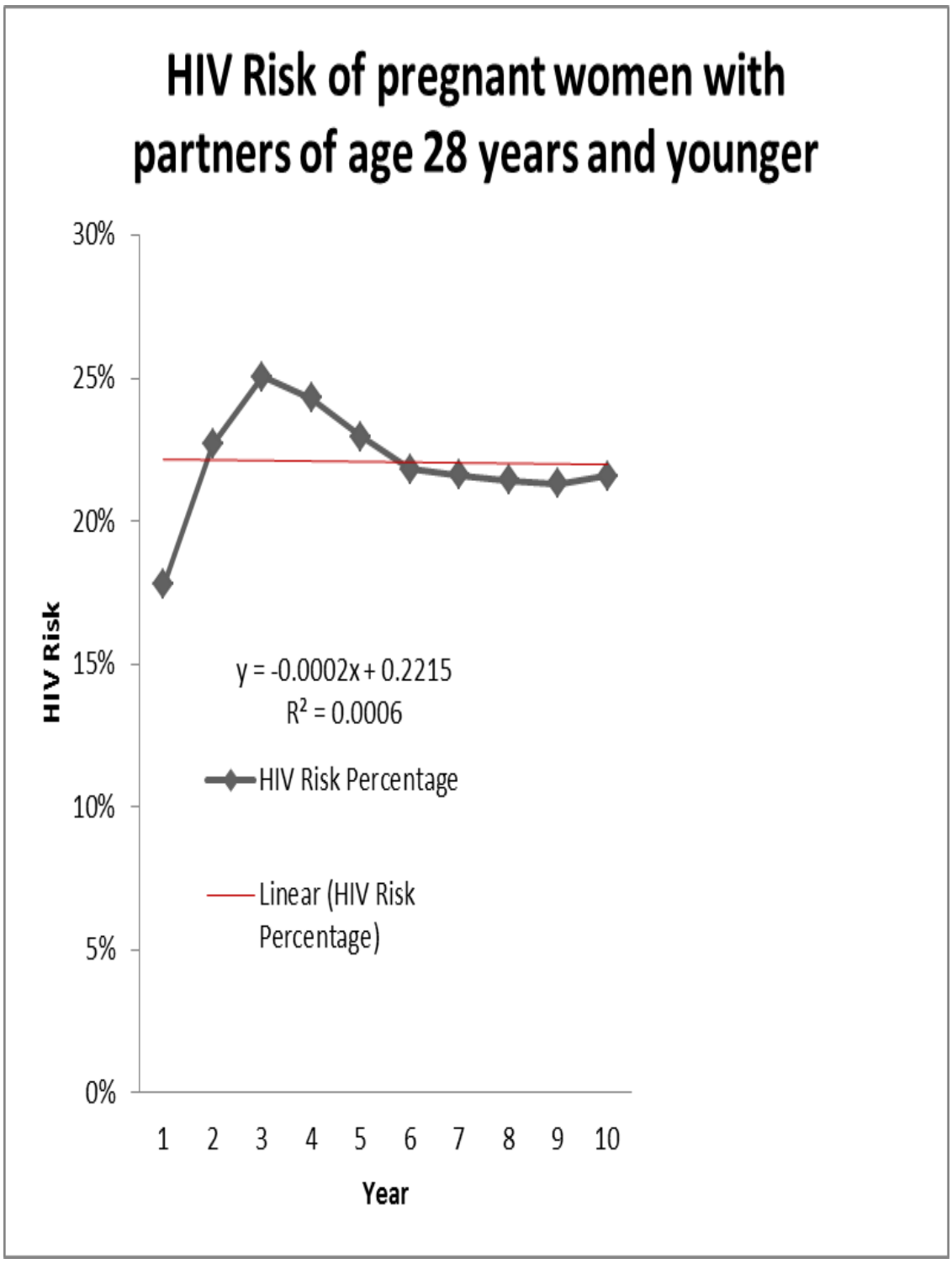


Figure 4-3: HIV risk to pregnant women with partners of ages 28 years and younger, 2001 to 2010.

FIGURE 4.3 shows a time series line graph of the risk of HIV of pregnant women with young partners. The graph shows that there is a slight decrease in the risk of HIV of pregnant women with partners 28 years and younger. The linear model is as follows:

$$Y = -0.0003\,x + 0.2227$$

The negative slope shows a downward trend, and that the risk over the ten-year period decreased by -0.0003.

The R-square show that the model only accounts for 0.23% of variability of the response data around the mean.

Table 4-4: HIV risk to pregnant women with partners of ages 28 years and older, 2001 to 2010.

Year	Coded Year	Total pregnant women	Total HIV positive pregnant women	HIV risk percentage
2001	1	9 288	2 215	23.85%
2002	2	8 158	2 449	30.02%
2003	3	1 036	295	28.47%
2004	4	7 851	2 670	34.01%
2005	5	6 472	2 303	35.58%
2006	6	16 169	5 597	35.23%
2007	7	16 616	5 994	36.07%
2008	8	16 411	6 006	36.60%
2009	9	16 003	5 908	36.92%
2010	10	15 014	5 764	38.39%

TABLE 4.4 shows that the risk of HIV for pregnant women with partners of ages 29 years and older increased over time.

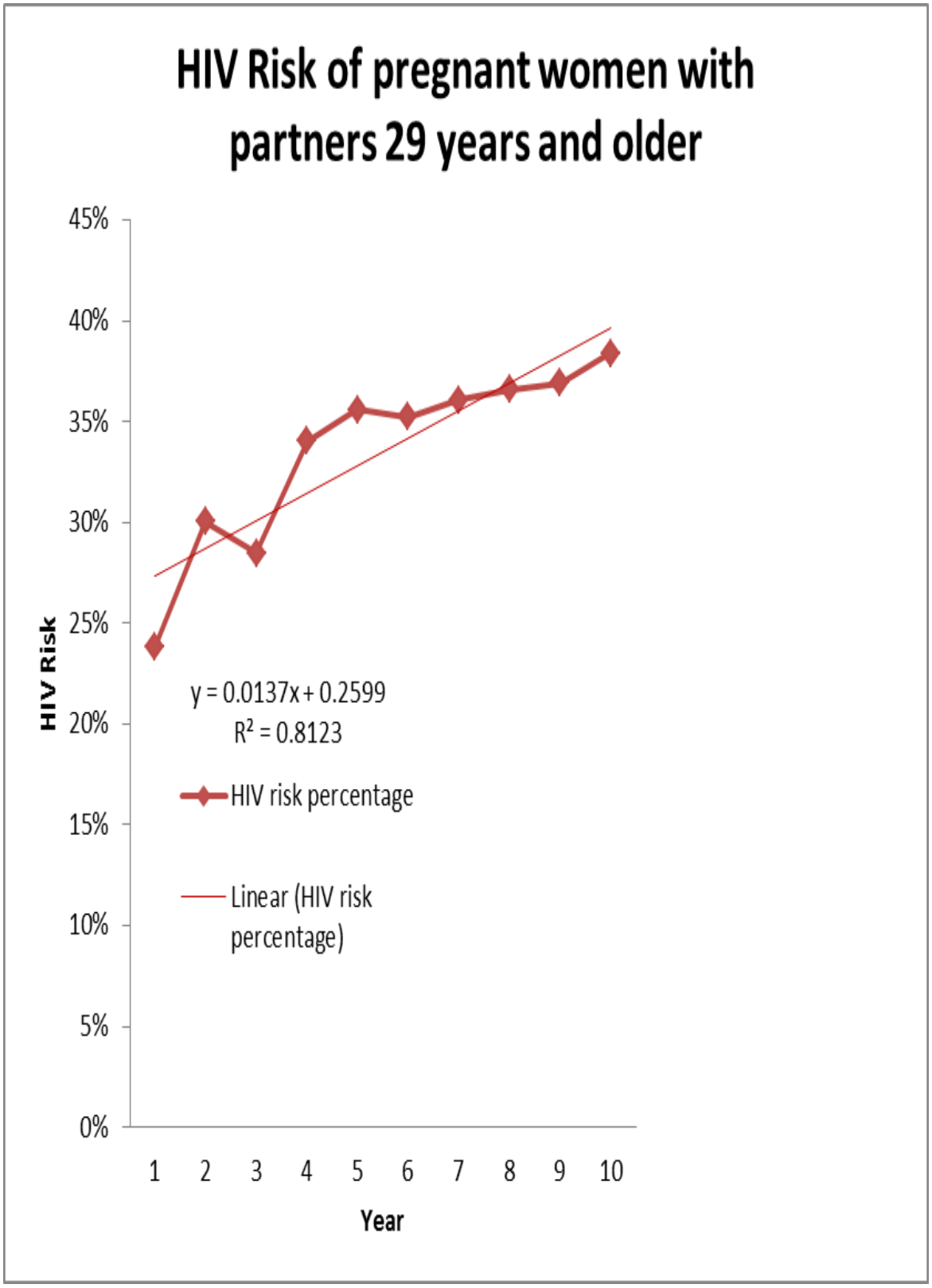


Figure 4-4: HIV risk to women with older partners, 2001 to 2010.

TABLE 4.4 and FIGURE 4.4 show the results of HIV risk for pregnant women with partners of ages 29 years and older from 2001 to 2010. The results show that there is an increase in HIV risk.

The linear model is as follows:

$$Y = 0.0137\,x + 0.2599$$

The positive slope shows an upward trend, which indicates an increase of the risk of HIV for pregnant women with partners of ages 29 years and older.

The R-square indicate that the model accounts for 79% of the variability of the response data around the mean.

Therefore the results found indicate that for the period 2001 to 2010, there is a higher HIV risk for pregnant women reported with partners of ages 29 years and older as compared to pregnant women with partners 28 years and younger.

4.2.3 HIV risk of pregnant women on gravidity

The demographic characteristic of gravidity signifies the number of pregnancies a woman has had, and in this study gravidity was divided into two levels, namely -1 and 1, where -1 represents women who are pregnant for the first time and 1 represents women who have had more than one pregnancy.

This section discusses the risk of HIV among women pregnant for the first time as well as the prevalence of HIV among women who have had one or more pregnancies. A conclusion is given at the end of this section.

TABLE 4.5 shows the results of pregnant women in their first pregnancy who tested positive for HIV. The TABLE 4.5 contains results of the total number of first-time pregnant women who were tested for HIV, the total number of first-time pregnant women who tested HIV positive and the percentage risk of HIV.

Table 4-5: HIV risk to pregnant women on gravidity (first pregnancy), 2001 to 2010.

Year	Coded Year	Gravidity(first pregnancy)	HIV risk on gravidity (first pregnancy)	HIV risk percentage
2001	1	4 797	1047	21.78%
2002	2	441	118	26.76%
2003	3	795	212	26.67%
2004	4	6 014	1 430	23.78%
2005	5	5 142	1 139	22.15%
2006	6	12 278	2 549	20.76%
2007	7	12 490	2 513	20.12%
2008	8	12 682	2 496	19.68%
2009	9	11 821	2 299	19.45%
2010	10	11 328	2 210	19.51%

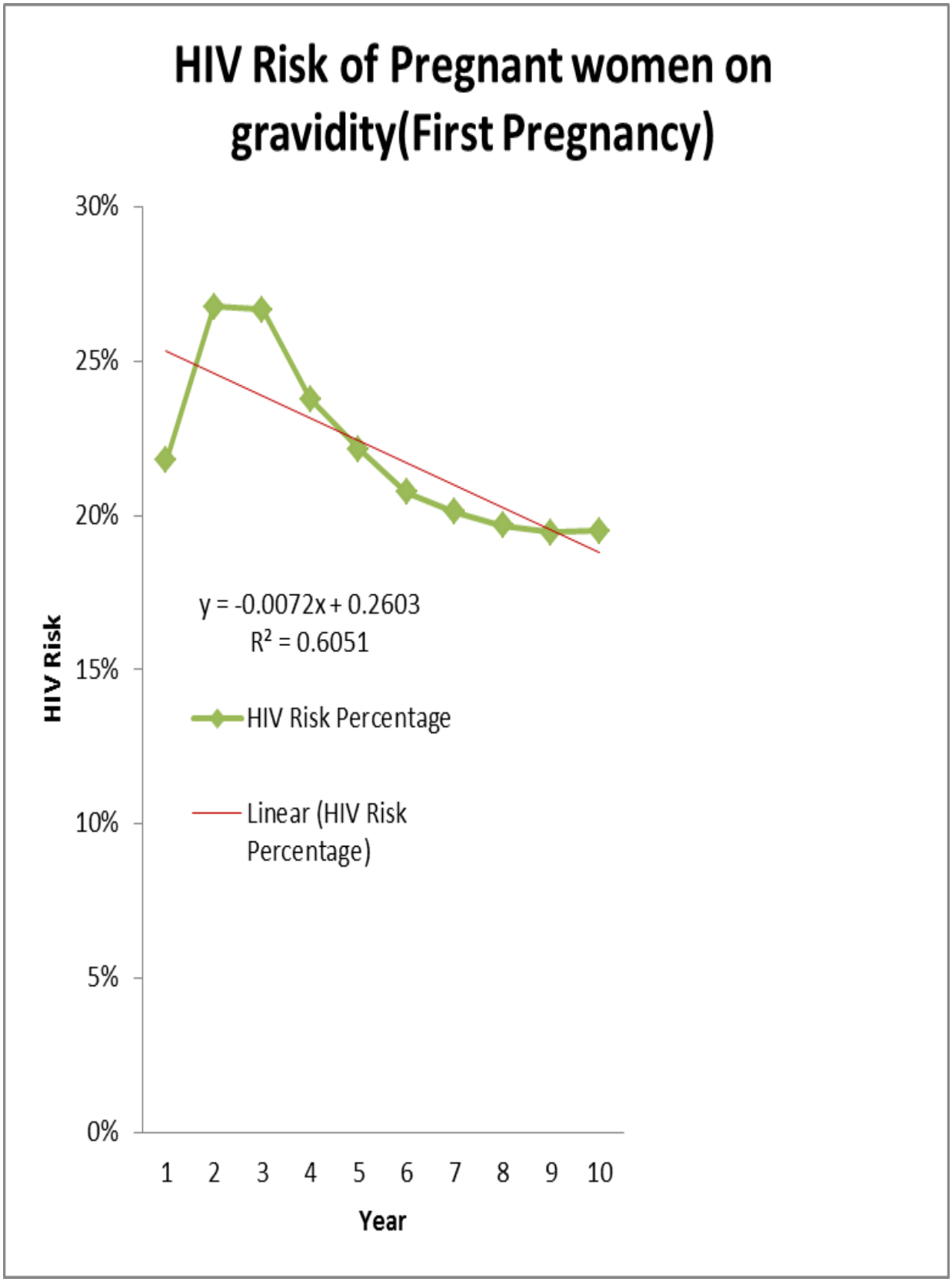


Figure 4-5: HIV risk to pregnant women on gravidity (first pregnancy), 2001 to 2010

TABLE 4.5 and FIGURE 4.5 show the risk among women who are pregnant for the first time. The results show that the HIV risk of first-time pregnant women has decreased over the years. There is therefore a downward trend, with 19% being the lowest percentage. The linear model is as follows:

$$Y = -0.00072x + 0.2603$$

The negative slope shows a downward trend in the risk of HIV over the ten-year period.

The R-square signifies that the model accounts for 60.51% of the variability of the response data around the mean.

TABLE 4.6 shows results of pregnant women with one or more pregnancies who tested positive for HIV, the total number of pregnant women who were tested for HIV and the HIV percentage risk.

Table 4-6: HIV risk to pregnant women on gravidity (one or more pregnancies), 2001 to 2010.

Year	Coded Year	Gravidity (one or more pregnancies)	HIV risk on gravidity (one or more pregnancies)	HIV risk percentage
2001	1	7 453	1 697	22.77%
2002	2	14 893	3 959	26.58%
2003	3	1 075	292	27.16%
2004	4	8 802	2 932	33.31%
2005	5	7 648	2 614	34.18%
2006	6	18 980	6440	33.93%
2007	7	19 991	6 909	34.56%
2008	8	19 889	6 974	35.06%
2009	9	19 322	6 835	35.37%
2010	10	18 784	6 813	36.27%

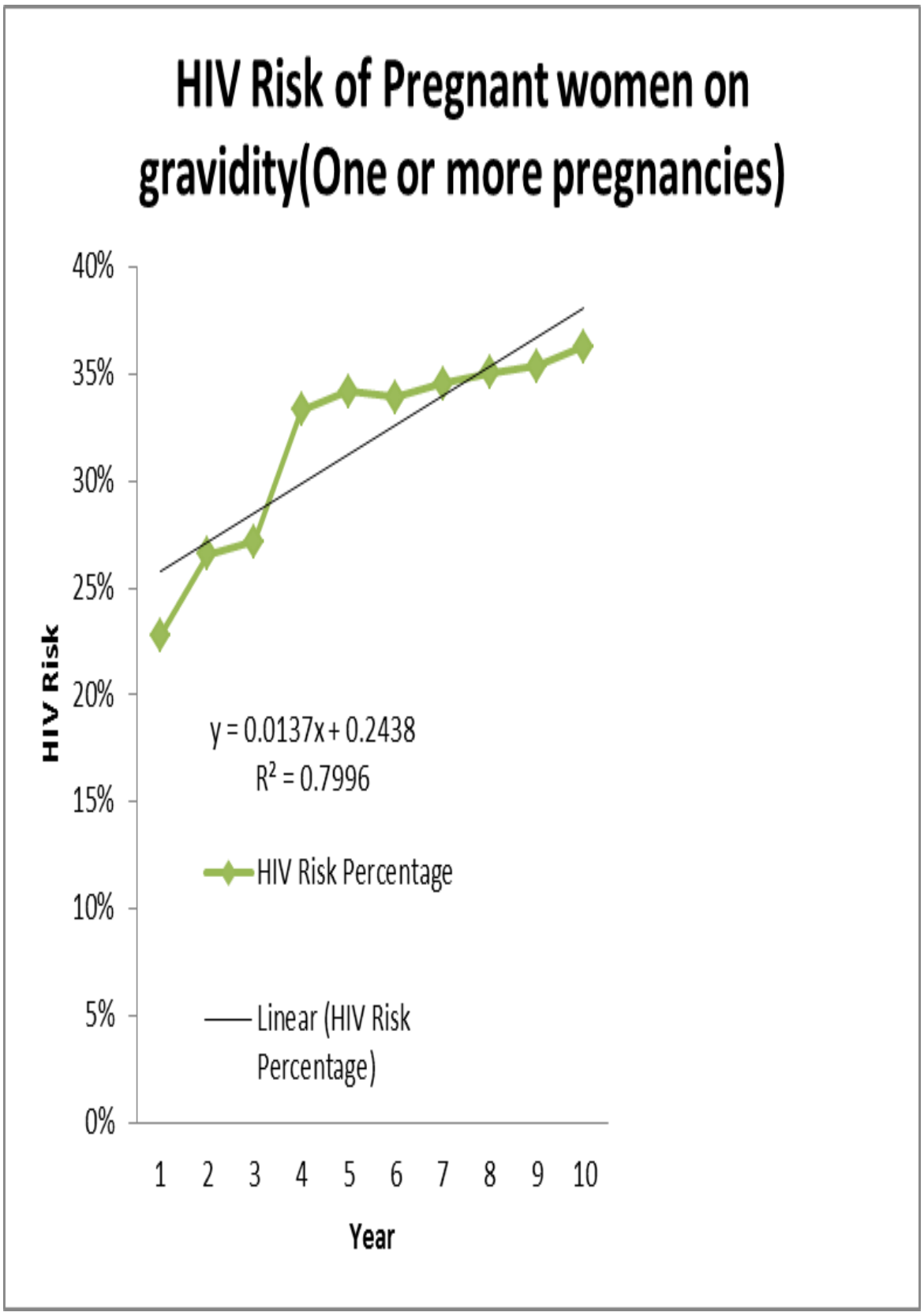


Figure 4-6: Control chart for HIV risk to pregnant women on gravidity (more than one pregnancy), 2001 to 2010

TABLE 4.6 and FIGURE 4.6 illustrate the risk of HIV among women who have had more than one pregnancy. The results show an increasing trend in the percentage of HIV risk of pregnant women with one or more than one children. The linear model produced is as follows:

$$y = 0.0137x + 0.2438$$

The positive slope shows an upward trend, which means that the risk of HIV for pregnant women who have had more than one pregnancy increased over the 10-year period.

The R-square show that the model explains 80% of the variability of the response data around the mean.

TABLE 4.6 and FIGURE 4.6 convey the fact that women who have had more than one pregnancy are at a higher risk of acquiring HIV as compared to women experiencing their first pregnancy.

4.2.4 HIV risk of pregnant women on parity

Parity signifies the number of children a woman has had, and in this study the demographic characteristic of parity was divided into two levels (- 1 and 1). Level -1 signifies women with no children yet, meaning that they are pregnant with their first child, and level 1 signifies women who already have one or more children.

The section below provides a closer look at both women with no children and women who have already had one or more children. A conclusion is given at the end of the section.

TABLE 4.7 shows the results of the total number of pregnant women tested for HIV, as well as the total number of pregnant women who tested HIV positive who had not yet had children. The percentage of the HIV risk is also shown.

Table 4-7: HIV risk to pregnant women on parity (no child), 2001 to 2010

Year	Coded Year	Parity (no child)	HIV risk on parity (no child)	HIV risk percentage
2001	1	5 449	1 203	22.08%
2002	2	6 134	1 259	22.16%
2003	3	839	232	27.65%
2004	4	6 027	1 446	23.99%
2005	5	5 465	1 267	23.18%
2006	6	13 198	2 897	21.95%
2007	7	13 419	2 853	21.26%
2008	8	13 782	2 906	21.09%
2009	9	12 716	2 656	20.89%
2010	10	12 211	2 542	20.82%

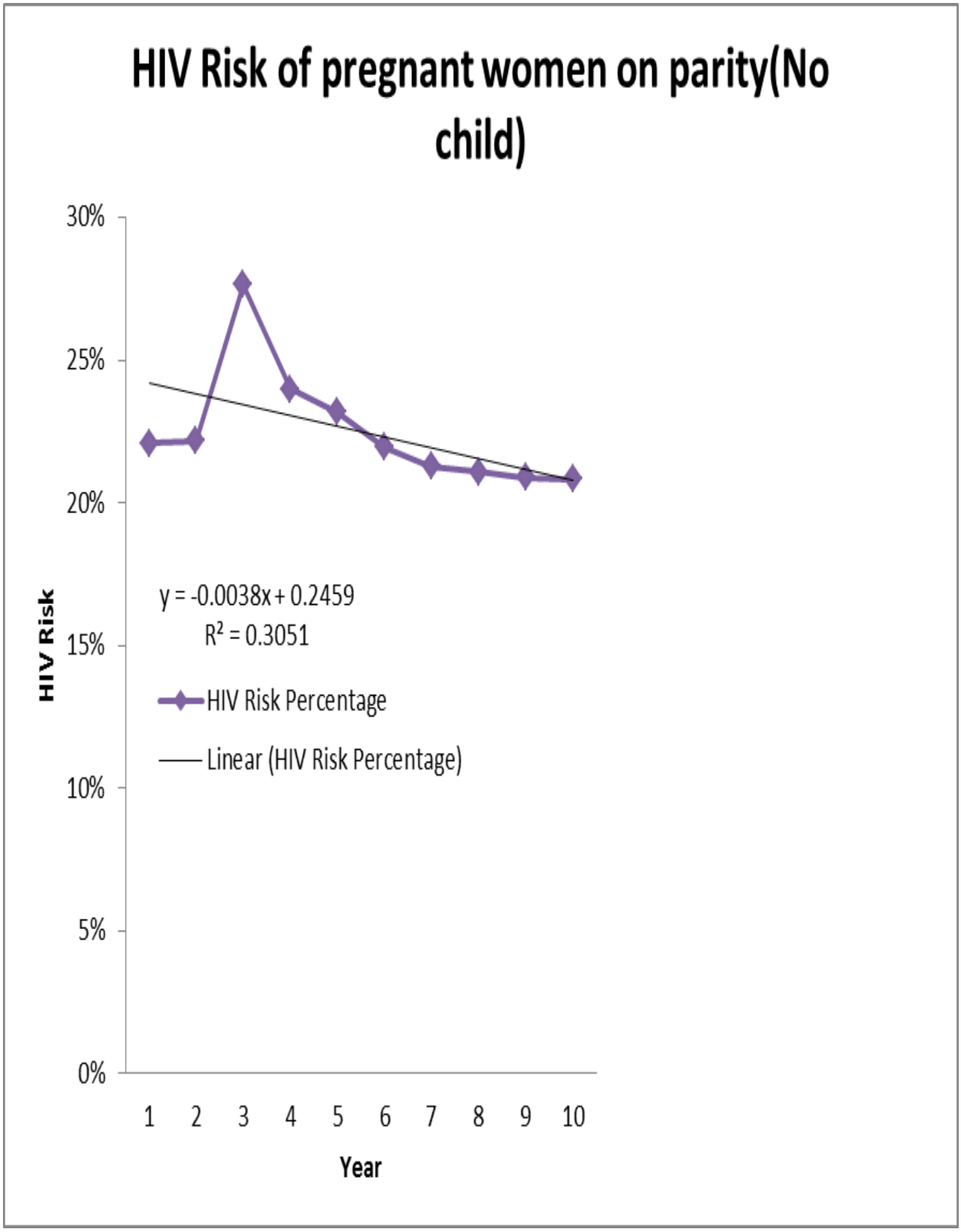


Figure 4-7: HIV risk to pregnant women on parity (no child), 2001 to 2010

TABLE 4.7 and FIGURE 4.7 show that there is a downward trend in the risk of HIV.

The linear model is as follows:

$$y = -0.0038x + 0.2459$$

The linear trend shows a negative slope, which indicates a decrease in the risk of HIV for pregnant women with no children over the ten-year period.

The R-square shows that the model accounts for 30.51% of the variability of the response data around the mean.

TABLE 4.8 shows the HIV risk of pregnant women on parity (one or more children) from 2001 to 2010.

Table 4-8: HIV risk to pregnant women on parity (one or more children), 2001 to 2010

Year	Coded Year	Parity (one or more children)	HIV risk on parity (one or more children)	HIV risk percentage
2001	1	6 801	1 539	22.63%
2002	2	9 199	2 718	29.55%
2003	3	1 031	272	26.38%
2004	4	8 789	2 916	33.18%
2005	5	7 325	2 486	33.94%
2006	6	18 060	6 092	33.73%
2007	7	19 062	6 569	34.46%
2008	8	18 789	6 564	34.94%
2009	9	18 427	6 478	35.15%
2010	10	17 901	6 481	36.20%

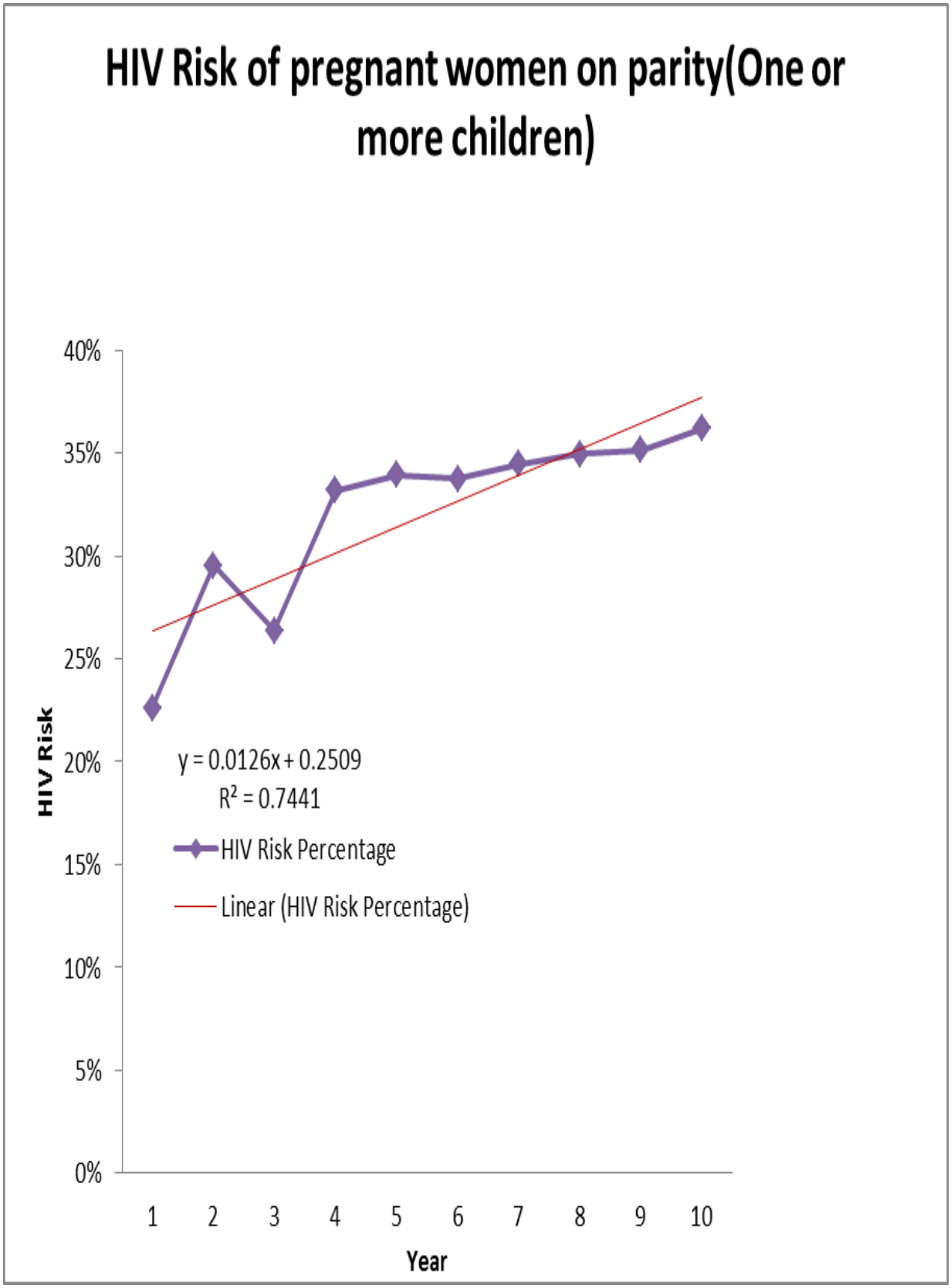


Figure 4-8: HIV risk to HIV-positive women on parity (one or more children), 2001 to 2010

TABLE 4.8 and FIGURE 4.8 show results for pregnant women with one or more children. The results show that there is an increasing trend in HIV risk.

The linear model generated is as follows:

$$y = 0.01236x + 0.2509$$

The linear model shows a positive slope, which indicates that there is an increase in the risk of HIV to pregnant women who have one or more children.

The model has an R-square of 0.7441, which indicates that the model explains 74% of the variability of the response data around the mean.

Therefore pregnant women with one or more children are at a higher risk of acquiring HIV as compared to those who have no children yet.

4.2.5 HIV risk to pregnant women on education

Education is one of the demographic characteristics of pregnant women that were captured with the aim of determining to what extend it affects the risk of pregnant women contracting HIV. The demographic factor of education was also divided into two levels, -1 and 1. The negative level (level -1) represents pregnant women with a primary school education as well as those who have no formal education at all. The positive level (level 1) represents pregnant women who have secondary/high school education as well as those with tertiary education.

In this section both levels are analysed to determine the effect that they have individually on the risk of HIV.

TABLE 4.9 shows the results of the total number of pregnant women with primary to no education tested for HIV and the total number who tested HIV positive. The HIV risk is also given.

Table 4-9: HIV risk to pregnant women with primary to no education, 2001 to 2010

Year	Coded Year	Education (primary to no education)	HIV risk on education (primary to no education)	HIV risk percentage
2001	1	341	95	27.86%
2002	2	6 482	1 501	23.16%
2003	3	20	3	15.00%
2004	4	304	89	29.28%
2005	5	228	65	28.51%
2006	6	523	165	31.55%
2007	7	1 054	329	31.21%
2008	8	363	137	37.74%
2009	9	567	167	29.45%
2010	10	475	164	34.53%

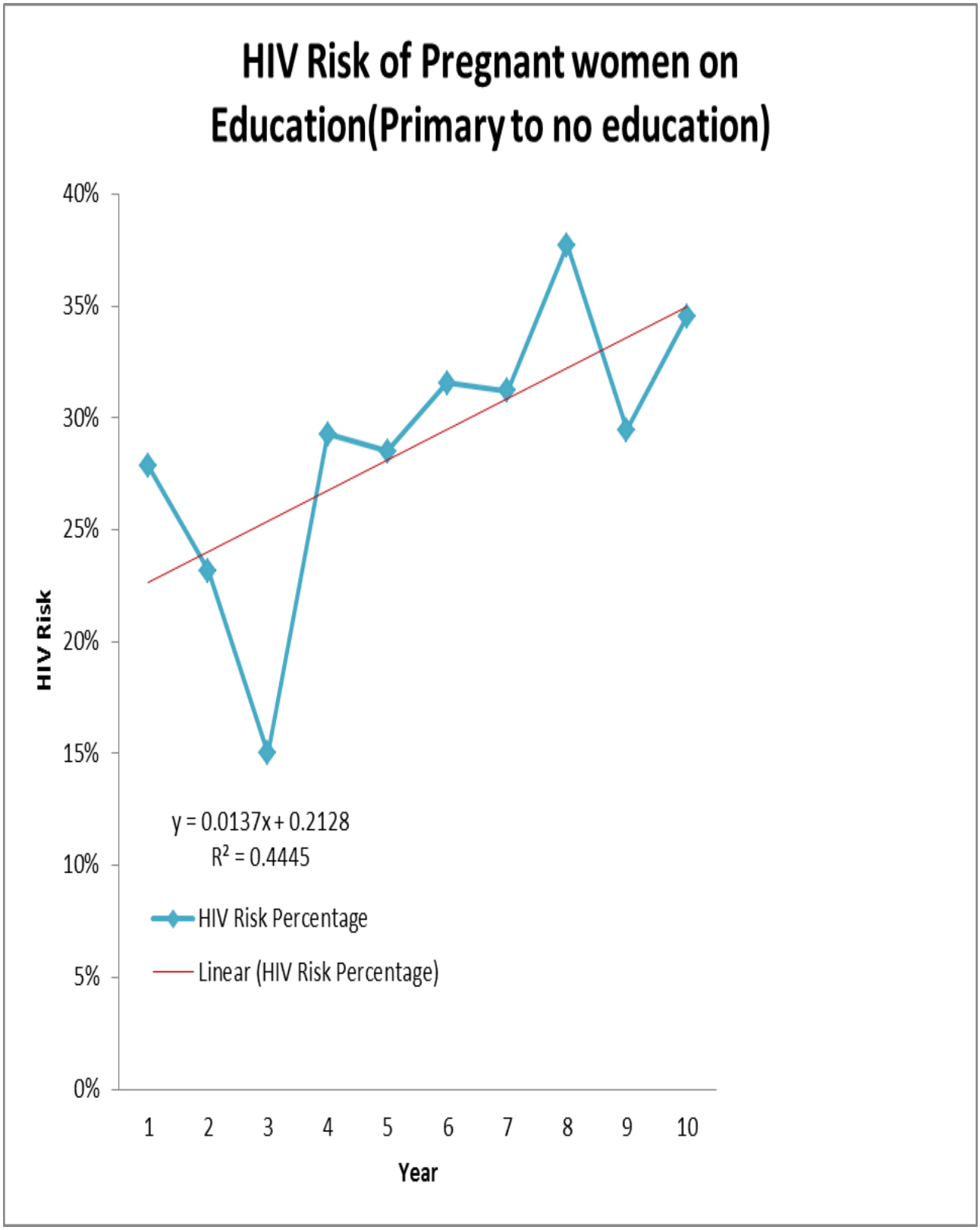


Figure 4-9: HIV risk to pregnant women with primary to no education, 2001 to 2010.

TABLE 4.9 and FIGURE 4.9 show the risk of HIV among pregnant women with primary to no education. Both TABLE 4.9 and FIGURE 4.9 show an increasing in the risk of HIV for pregnant women with primary to no education.

The linear model generated is as follows:

$$y = 0.0137x + 0.2128$$

The linear model shows a positive slope which, like Figure 4.9, shows an upward or increasing trend in the risk of HIV to pregnant women with primary to no educational background.

The model has an R-square of 0.4445, which shows that the model accounts for 44% of the variability of the response data around the mean.

TABLE 4.10 shows the results of the total number of pregnant women with secondary to tertiary education who were tested for HIV and the total number who tested HIV positive. The HIV risk percentage is also given.

Table 4-10: HIV risk to pregnant women with secondary to tertiary education, 2001 to 2010.

Year	Coded Year	Education (secondary to tertiary)	HIV risk to pregnant women with education (secondary to tertiary)	HIV risk percentage
2001	1	11 909	2 647	22.23%
2002	2	8 851	2 576	29.10%
2003	3	1 850	501	27.08%
2004	4	14 512	4273	29.44%
2005	5	12 562	3 688	29.36%
2006	6	30 735	8 824	28.71%
2007	7	31 427	9 093	28.93%
2008	8	32 208	9 333	28.98%
2009	9	30 576	8 967	29.33%
2010	10	29 637	8 859	29.89%

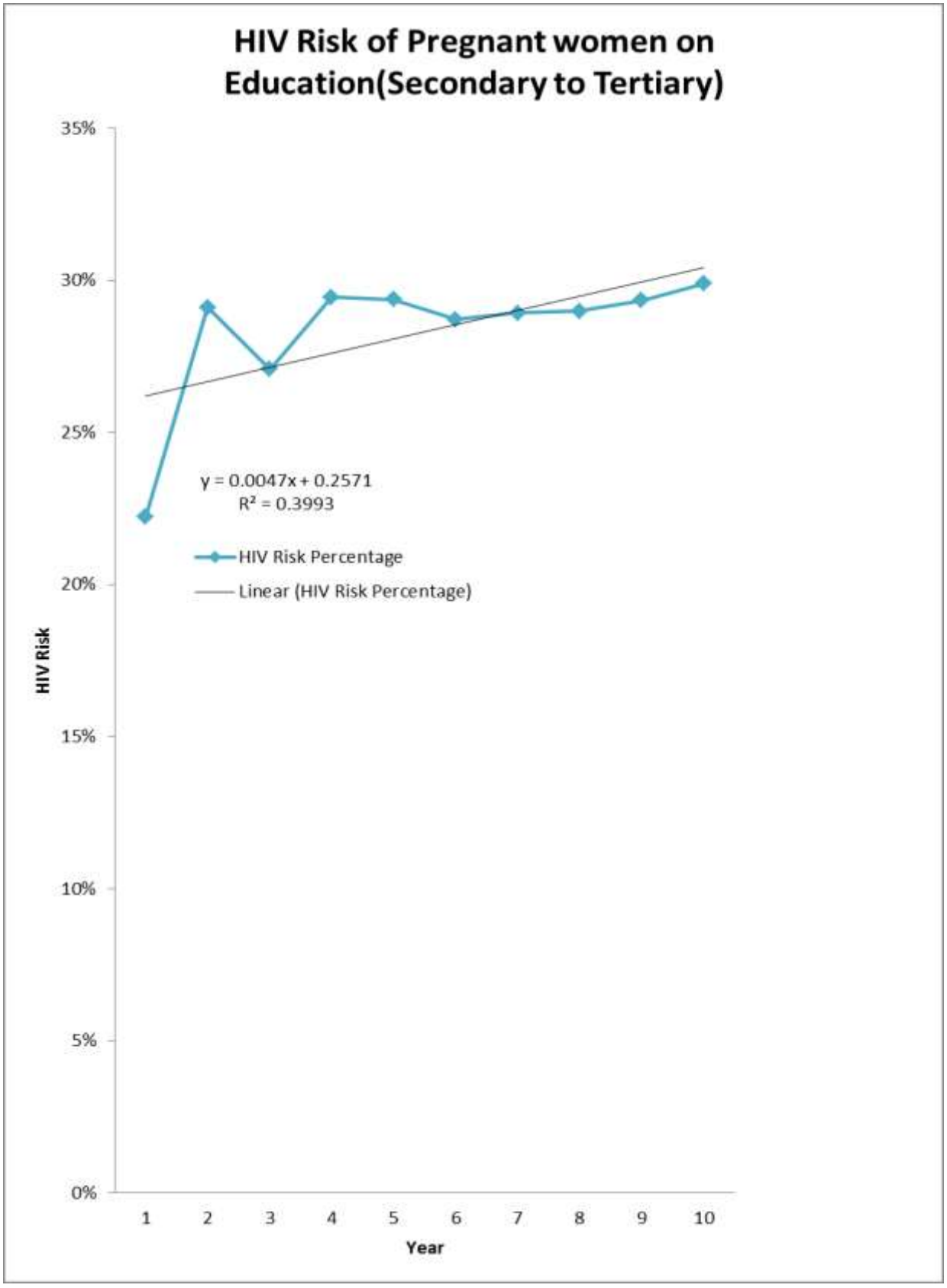


Figure 4-10: HIV risk to pregnant women with secondary to tertiary education, 2001 to 2010.

TABLE 4.10 and FIGURE 4.10 show the risk of HIV among pregnant women with secondary to tertiary education. When looking at the results above, the risk of acquiring HIV to pregnant women with secondary to tertiary education shows an increasing trend. The risk was at its lowest in 2001 at 21% and at its highest in 2010 at 30%, and the risk is above average from 2004.

The linear model generated is as follows:

$$y = 0.0047x + 0.2571$$

The linear model has a positive slope, which shows that the risk of HIV to pregnant women over the ten-year period experienced an increase.

The model has an R-square of 0.3993, which shows that the model accounts for 40% of the variability of the response data around the mean.

4.2.6 HIV risk to pregnant women on syphilis

This section investigates the risk of HIV to pregnant women who tested positive for syphilis and HIV, as well as the risk of HIV to pregnant women who tested negative for syphilis but tested positive for HIV.

TABLE 4.11 shows the results of the total number of syphilis-positive pregnant women tested for HIV and the total number who tested positive for HIV. The risk percentage for HIV is also given.

Table 4-11: HIV risk to syphilis-positive pregnant women, 2001 to 2010

Year	Coded Year	Total syphilis-positive pregnant women	Total number of HIV-positive syphilis-positive pregnant women	HIV risk percentage
2001	1	400	109	27.25%
2002	2	514	220	42.80%
2003	3	70	25	35.71%
2004	4	340	138	40.59%
2005	5	313	116	37.00%

2006	6	588	216	36.73%
2007	7	892	335	37.56%
2008	8	622	231	37.16%
2009	9	600	245	40.83%
2010	10	447	196	43.85%

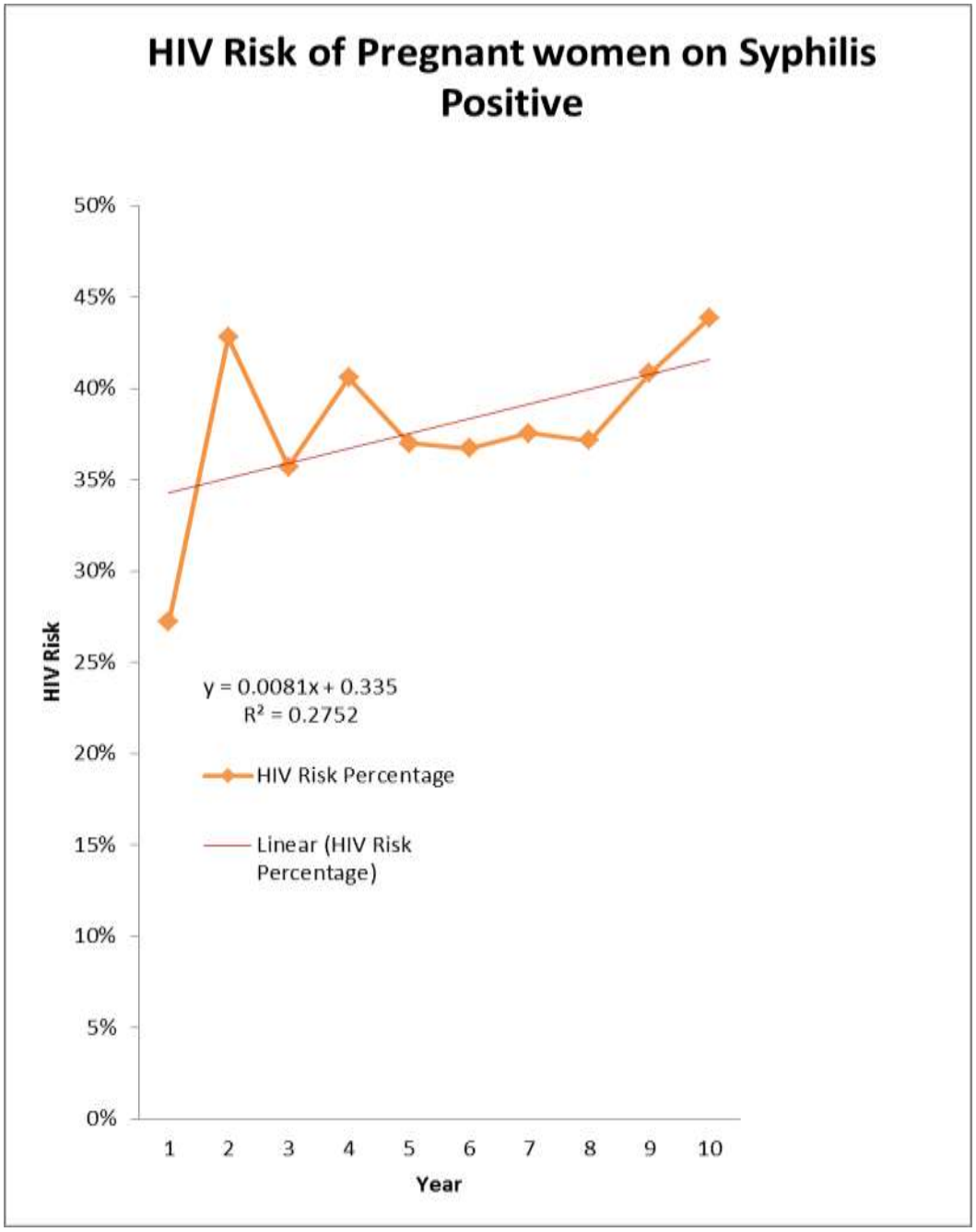


Figure 4-11: HIV risk to pregnant women on syphilis positive, 2001 – 2010

As mentioned previously, syphilis is a sexually transmitted disease, which is known to increase the risk of a person acquiring HIV.

FIGURE 4.11 shows the results of pregnant women who tested positive for both HIV and syphilis. The results show an increase in the risk of HIV.

The linear model generated is as follows:

$$y = 0.0081x + 0.335$$

The linear model shows a positive slope, which means that the risk for syphilis-positive pregnant women of acquiring HIV increased over the ten years.

The model has an R-square of 0.2752. This shows that the model accounts for 47.52% of the variability of the response data around the mean.

TABLE 4.12 shows results of pregnant women who tested positive for HIV, but tested negative for syphilis, and the HIV risk percentage.

Table 4-12: HIV risk to syphilis-negative and HIV-positive pregnant women, 2001-2010

Year	Coded Year	Total syphilis-negative pregnant women	Total number of HIV-positive, syphilis-negative pregnant women	HIV risk percentage
2001	1	11 850	2 633	22.22%
2002	2	14 819	3 857	26.03%
2003	3	1 800	479	26.61%
2004	4	14 476	4 228	29.18%
2005	5	12 477	3 637	29.15%
2006	6	30 670	8 773	28.60%
2007	7	31 589	9 087	28.77%
2008	8	31 949	9 239	28.92%

2009	9	30 543	8 889	29.10%
2010	10	29 665	8 827	29.76%

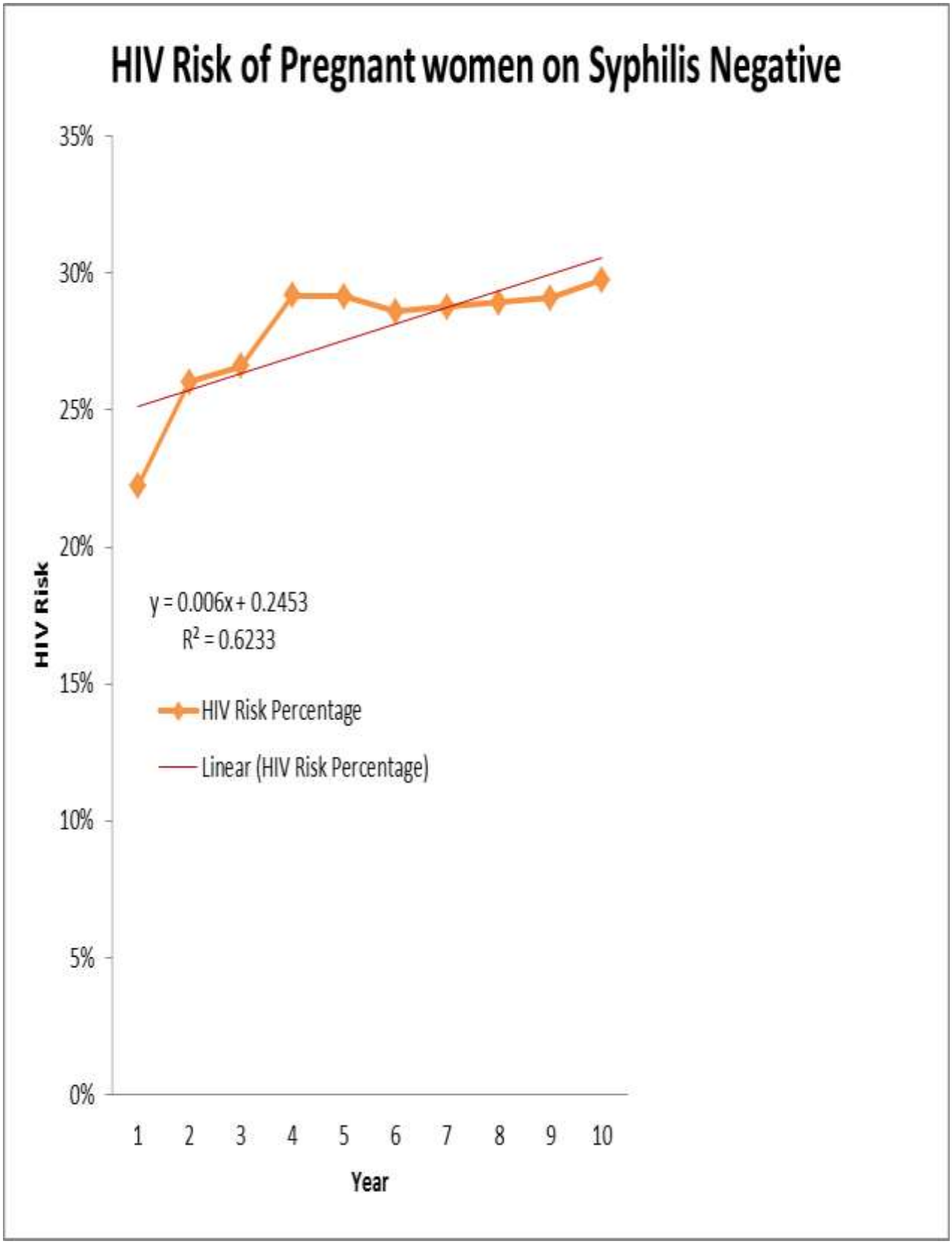


Figure 4-12: HIV risk to pregnant women on syphilis negative, 2001 to 2010.

TABLE 4.12 and FIGURE 4.12 show the risk of HIV to women who tested negative for syphilis. The results show an increase in the risk of HIV.

The linear model generated is as follows:

$$y = 0.006x + 0.2453$$

The linear model has a positive gradient, therefore showing that the risk of HIV to syphilis-negative pregnant women increased over the ten-year period.

The model has an R-square of 0.6233, meaning the model only accounts for 62% of the variability of the response data around the mean.

The results found show that syphilis still contributes to the risk of a pregnant woman acquiring HIV. Pregnant women who tested positive for syphilis are still at a higher risk of HIV than pregnant women who tested negative for syphilis.

4.3 DIFFERENTIAL EFFECTS

The section below seeks to analyse the differential effects of all the demographical characteristics of the pregnant women, as well as determine trends within the demographical characteristics for ten years. A differential effect is the difference in HIV risk between the high level (1) and the low level (-1) of a demographical variable.

4.3.1 Mother's age differential effect

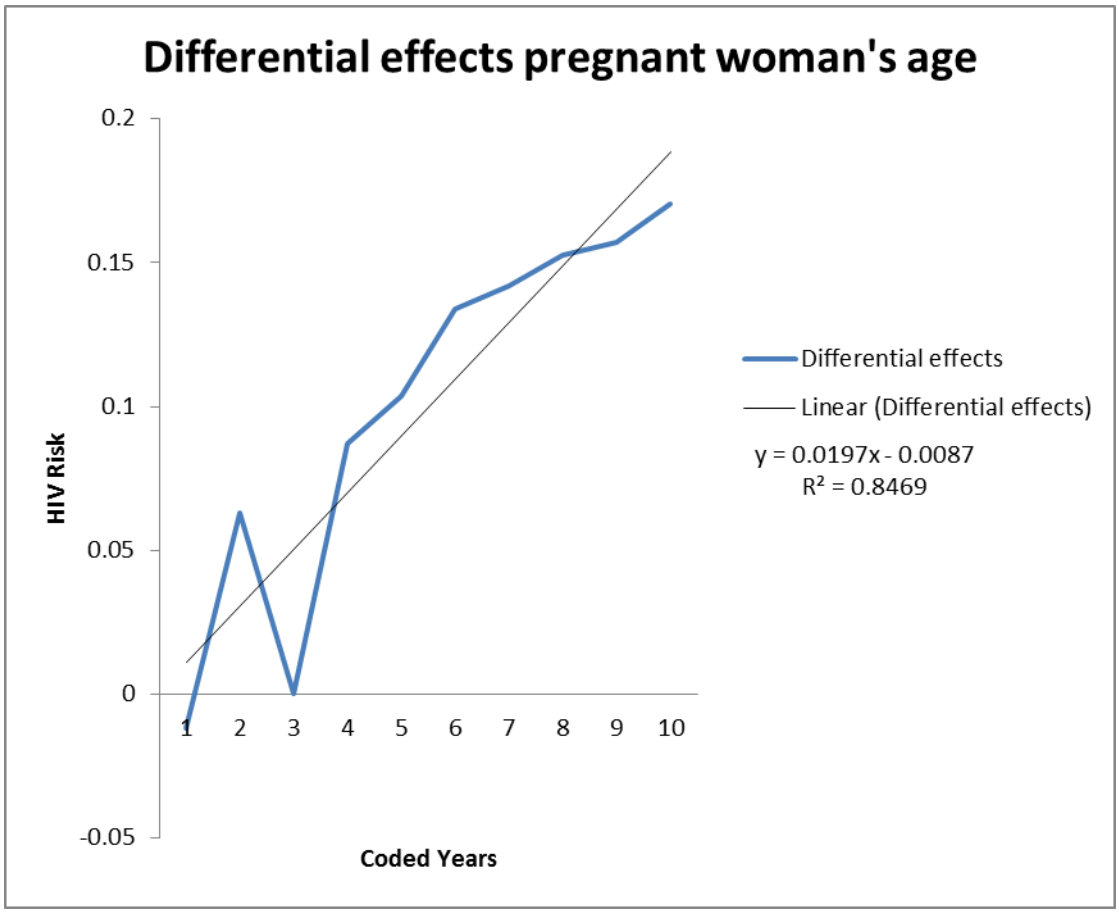


Figure 4-13: Differential effect on mother's age 2001 to 2010.

The differential effect on mother's age shows an increase over time.

The percentile difference in the HIV risk between older pregnant women compared to younger pregnant women has increased over time.

4.3.2 Pregnant women's partners age differential effects

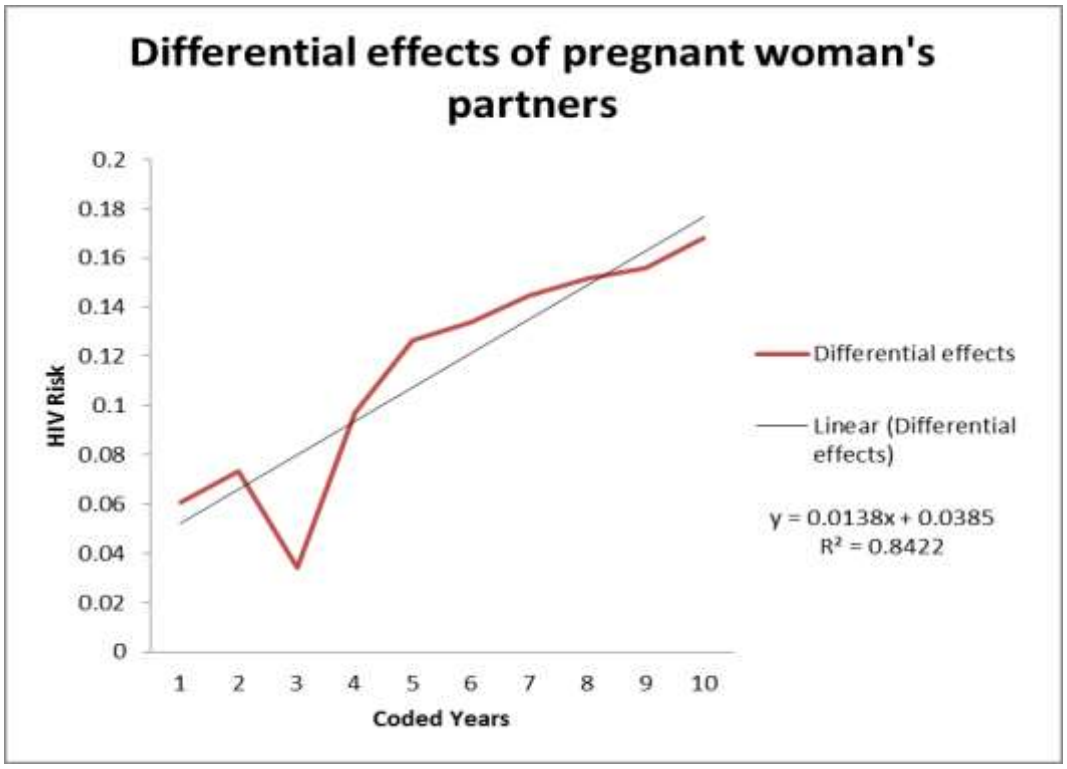


Figure 4-14: Pregnant women's partners age differential effects.

The differential effect of the pregnant woman's partner's age shows an increase in the risk of HIV over time.

The percentile difference in the HIV risk between pregnant woman's partners age has increased over time.

4.3.3 Pregnant woman's gravidity differential effects

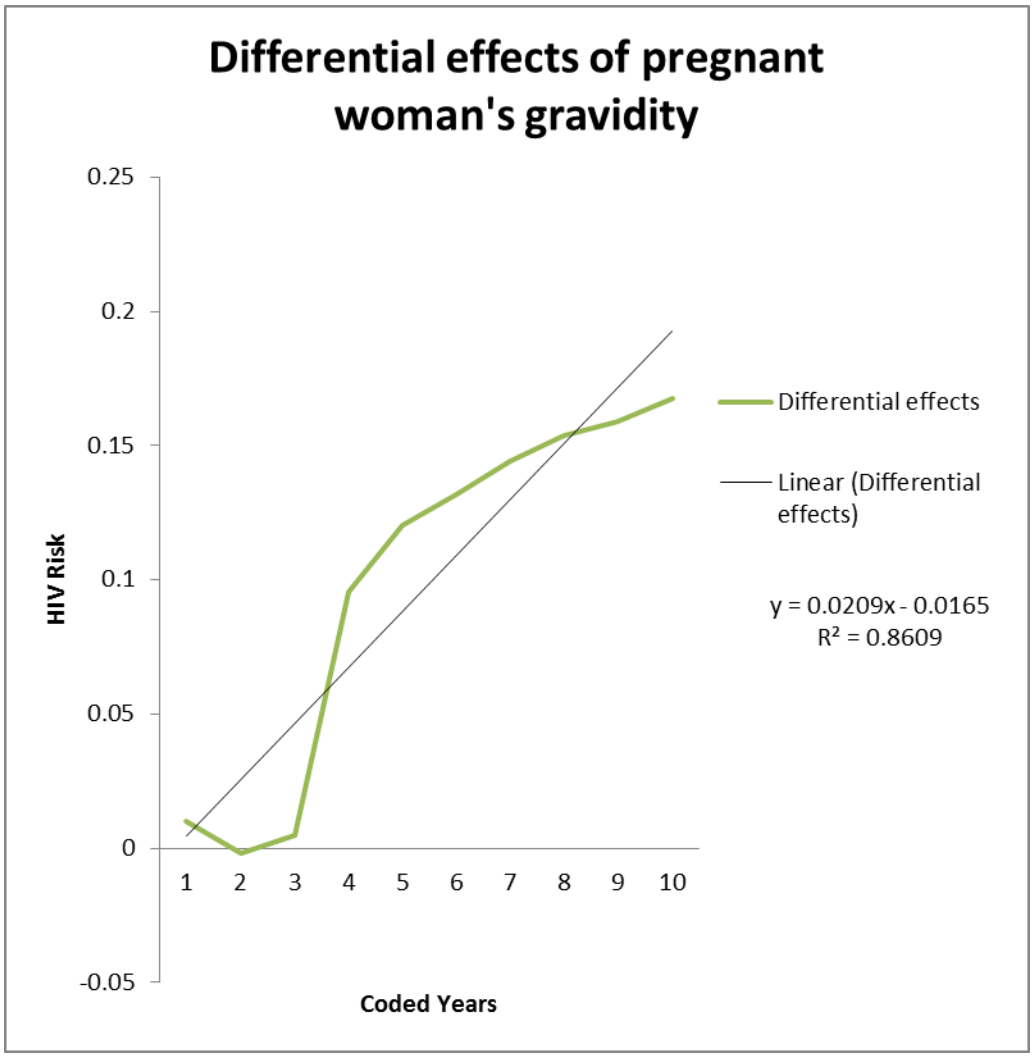


Figure 4-15: Differential effects of pregnant woman's gravidity.

The differential effects of the pregnant woman's gravidity show an increase on the risk of HIV over time.

The percentile difference in the HIV risk of pregnant woman's gravidity has increased over time.

4.3.4 Pregnant woman's parity differential effects

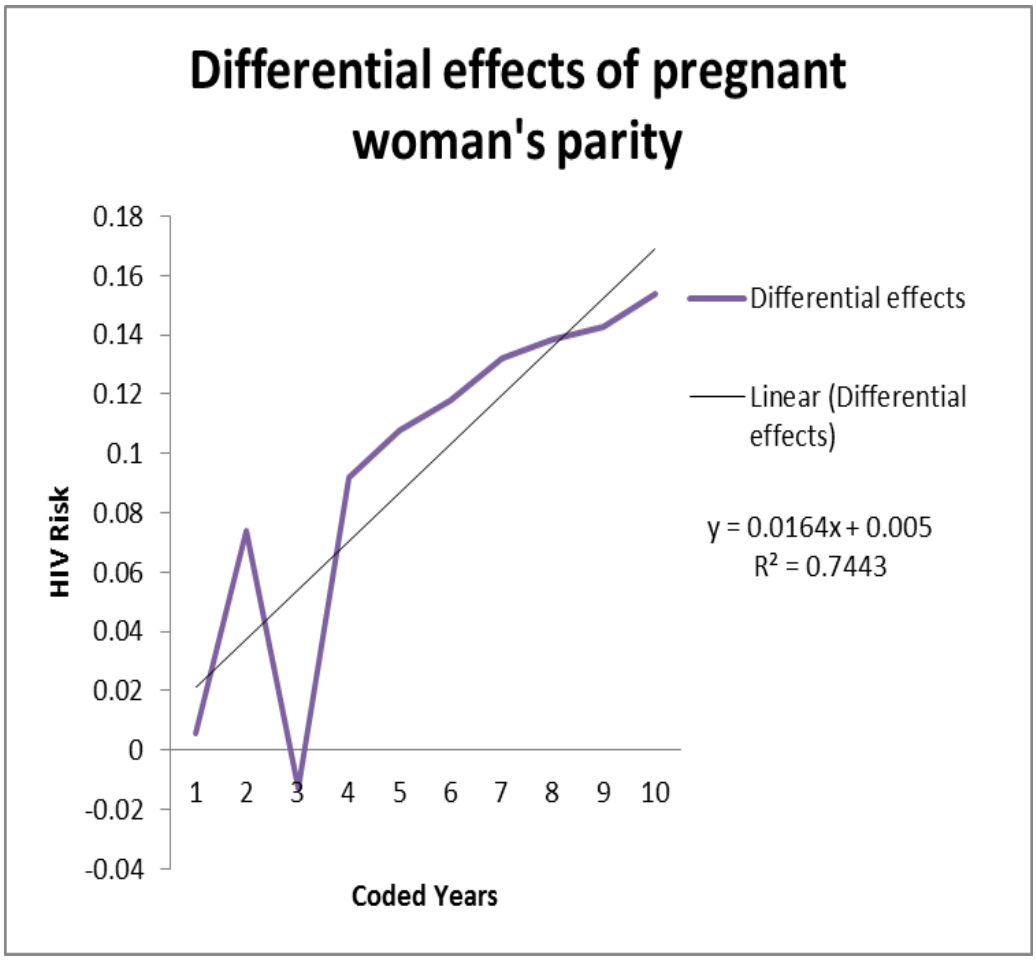


Figure 4-16: Differential effects of pregnant woman's parity

The differential effects of the pregnant woman's parity show an increase on the risk of HIV over time.

The percentile difference in HIV risk of the pregnant woman's gravidity has increased over time.

4.3.5 Pregnant woman's education differential effects

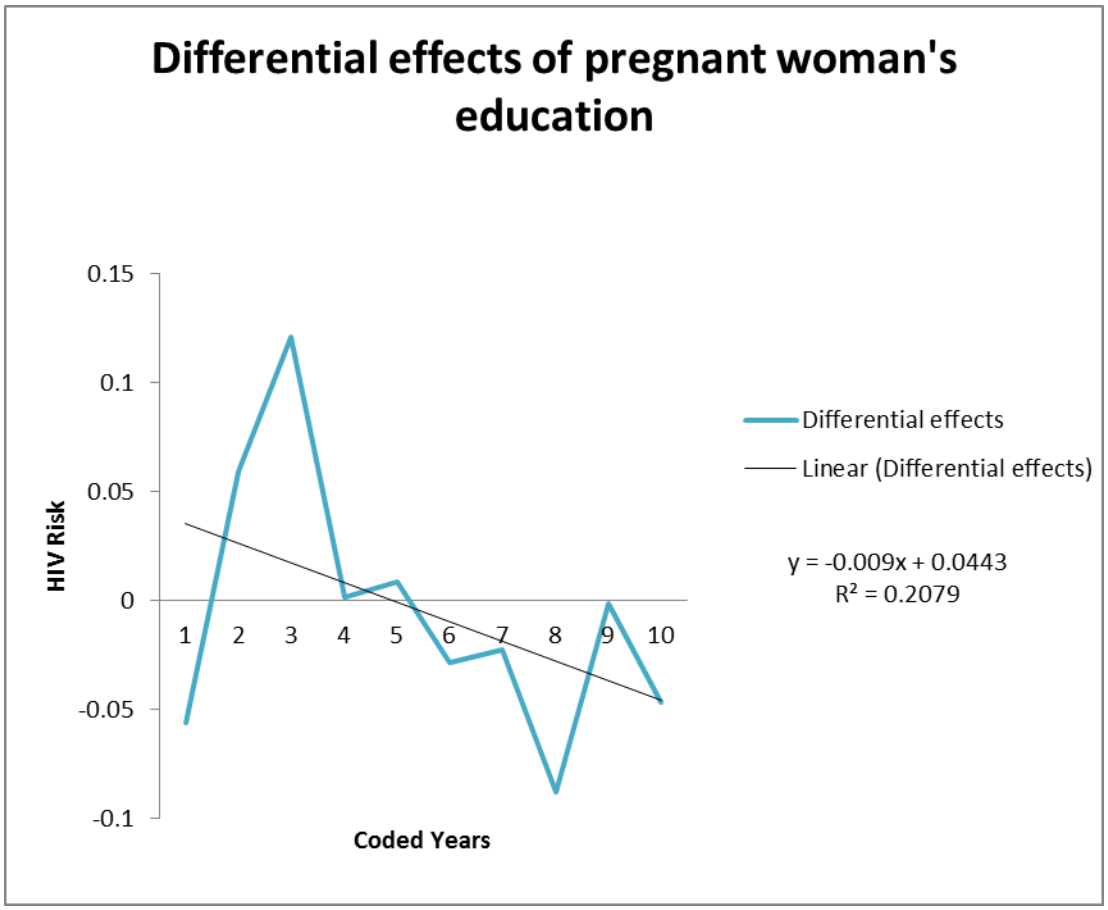


Figure 4-17: Differential effects of pregnant woman's education.

The differential effects of the pregnant woman's level of education show a decrease in the risk of HIV over time.

The percentile difference in HIV risk between the pregnant woman's level of education has decreased over time.

4.3.6 Pregnant woman's syphilis status differential effects

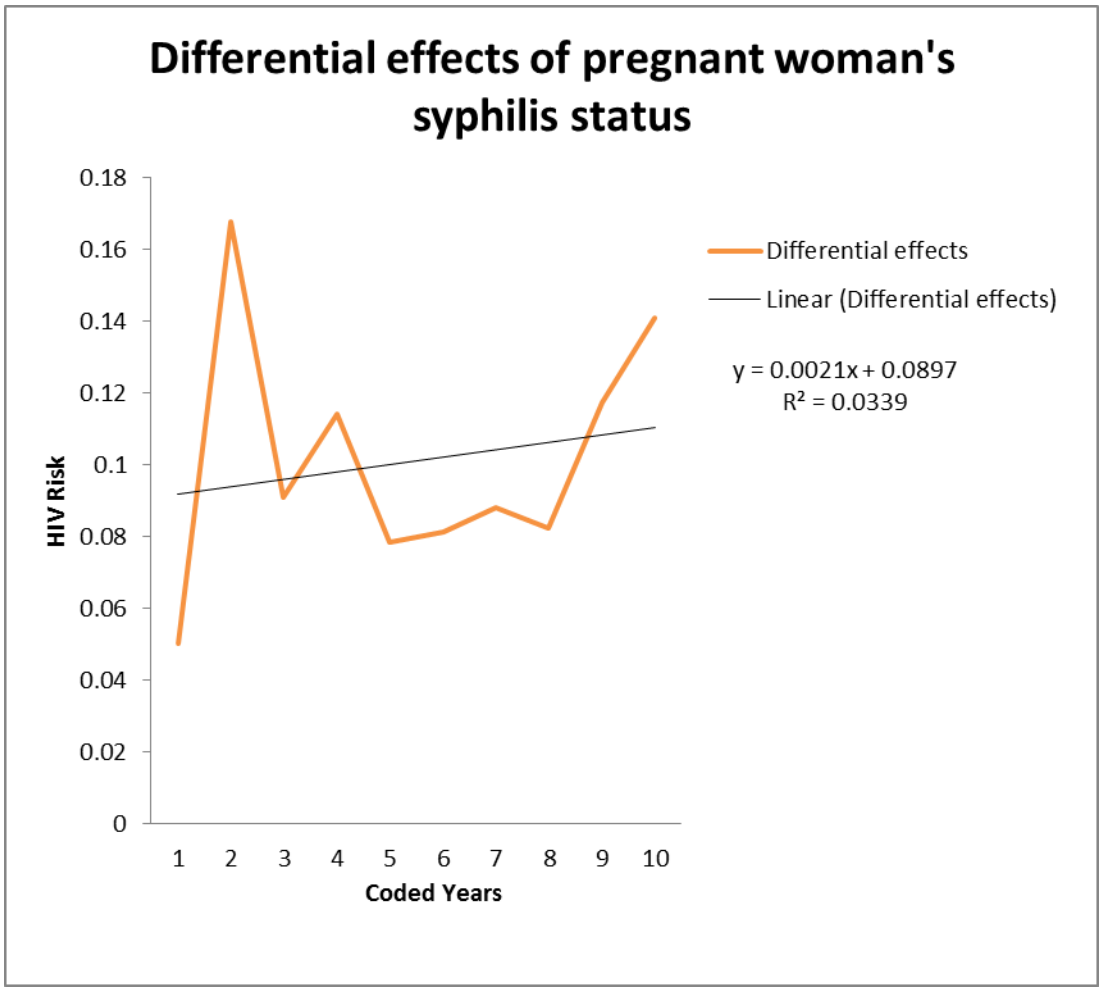


Figure 4-18: Differential effects of pregnant woman's syphilis status.

The differential effects of the pregnant woman's syphilis status show a slight increase in the risk of HIV over time.

The percentile difference in the HIV risk between the pregnant woman's syphilis statuses has increased over time.

4.4 CHAPTER CONCLUSION

In this chapter the results revealed the fact that over the ten-year period pregnant women of ages 25 and older were at a higher risk of acquiring HIV compared to pregnant women of ages 24 and younger.

The results also showed that pregnant women who tested positive for syphilis were at a higher risk of acquiring HIV compared to pregnant women who tested negative for syphilis.

The demographic factors parity and gravidity revealed the same results, namely that pregnant women who previously had no children and were experiencing their first pregnancy were at a lesser risk of acquiring HIV compared to pregnant women who already had children and had had more than one pregnancy. The time trend analysis run on parity and gravidity reveals that both parity and gravidity showed an upward trend over time, and signifies that parity and gravidity has a positive effect on the risk of HIV to pregnant women.

A comparison between pregnant women who had no children and those who had children revealed that pregnant women who had children were at a higher risk of acquiring HIV compared to those with no children.

A comparison between pregnant women who had been pregnant before and those experiencing their first pregnancy revealed that the pregnant women who had had more than one pregnancy were at a higher risk.

Analysis of the demographic factor of education revealed that pregnant women who had secondary to tertiary education were at a higher risk of HIV compared to those with primary to no education.

A comparison of pregnant women with partners of ages 28 years and younger and pregnant women with partners of ages 29 years and older revealed that the HIV risk to the latter was much higher than that of the former from 2001 to 2010.

The year 2003 seems to be an outlier due to the odd results generated from its data.

The objectives of this chapter were achieved the demographic characteristics of pregnant women using time series regression were analysed and interpreted.

The next chapter looks at the development of the two-level full factorial models.

CHAPTER 5: DEVELOPMENT OF A TWO-LEVEL FULL FACTORIAL MODEL

5.1 INTRODUCTION

CHAPTER 4 discusses time-series linear modelling, and takes a closer look at the pregnant women's demographic characteristics over the ten-year period. It also presents linear models for all six of the pregnant woman's demographics, which show a change over time.

The objective of this study is to develop two-level full factorial models to analyse HIV data. To achieve this objective, the Design of Experiments methodology was formulated in the study discussed in Chapter two, which explained that a two-level full factorial model allows the analysis of multiple factors simultaneously.

The objective of this chapter is to demonstrate understanding of the actual development of the two-level full factorial models, as well as to provide the two-level full factorial matrix. This chapter also aims to answer the research question stated in Chapter 1, namely whether the model remained stationary over the ten-year period.

In order to achieve its objective, this chapter is divided into the following sections: development of a two-level full factorial model (Section 5.2), model analysis (Section 5.3) finally conclusions (Section 5.4).

5.2 DEVELOPMENT OF A TWO-LEVEL FULL FACTORIAL MODEL

A factorial design is a method used in DOE which Morris explains as a factorial treatment structure where the effect of many different factors or treatments are tested by varying them simultaneously (Morris, 2011). The use of a full factorial design requires that an experimental run be performed at every combination of the factor level (JMP, 2014).

A two-level full factorial design has a sample that is to the power of two, which is described as 2^k, where k is the number of factors (Jaynes, 2013). In this study six factors were considered, namely: pregnant women's age, father's age, parity,

gravidity, level of mother's education and syphilis, and each was tested individually and simultaneously to determine their effect on the risk of HIV.

The study used coded HIV antenatal data, as this was the only data set made available to the researcher. The coded antenatal data was split in two levels based on the demographic characteristics of the pregnant women. The demographic characteristics of the pregnant women, which are the factors in question, are as follows:

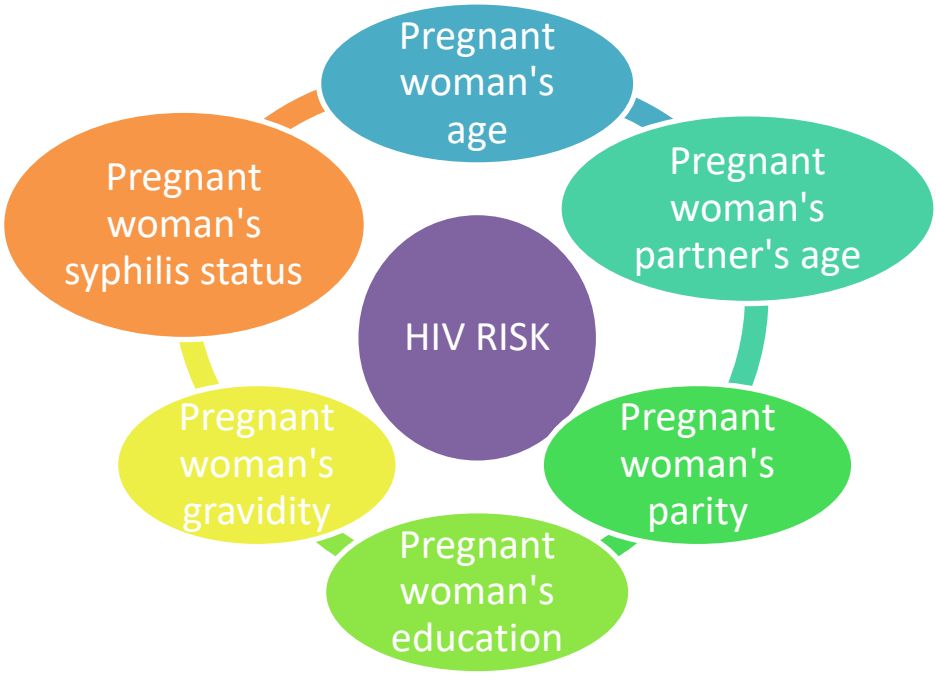


Figure 5-1: Demographic factors of pregnant women

FIGURE 5.1 gives the demographic characteristics of the pregnant women. The circle in the centre shows that all the factors may influence the risk of HIV one way or another.

5.2.1 Two-level full factorial design points

The section below shows how the two-levels in this design were determined based on all the factors and it also shows the complete two-level full factorial matrix developed from the six factors.

Table 5-1: Factors and levels table

	Levels	
Factors	**-1**	**1**
Mother's age	<= 24	>24
Father's age(Pregnant woman's partner)	<= 28	>28
Education (grades)	Primary to no education	Secondary and tertiary
Gravidity (number of pregnancies)	1	>1
Parity (number of children)	0	>1
Syphilis	0	1

TABLE 5.1 shows the six factors as well as the levels, which are described as follows: pregnant women from ages 24 and less are represented by the negative 1 (level -1) and pregnant woman of ages 25 and older are represented by the positive 1 (level 1).

Fathers of age 28 and less are represented by the negative one (level -1) and fathers of ages 29 and above are represented by positive 1 (level 1).

The educational level that was captured is for pregnant women, and therefore negative 1 means that the woman has primary school education to no education, and the positive 1 means that the woman has secondary to tertiary level education.

Gravidity, which is the number of pregnancies the women have had is represented as follows: negative 1 means that it is the woman's first pregnancy and positive 1 means that the woman has had more than one pregnancy.

Parity, which is the number of children the woman has had, is represented as follows: negative 1 means that the pregnant woman has no children yet, and positive one means the pregnant woman has more than one child.

Finally, syphilis is represented as follows: negative 1 means that the pregnant woman tested negative for syphilis and positive 1 means that the pregnant woman tested positive for syphilis.

TABLE 5.1 assisted in constructing the two-level full factorial design matrix. A two-level full factorial model is denoted as 2^k, where *k* is the number of factors. In this study there are six factors, which therefore results in 64 runs as illustrated in TABLE 5.2.

Table 5-2: Two-level full factorial matrix

Run	Mother's age	Father's age	Education	Gravidity	Parity	Syphilis
1.	-1	-1	-1	-1	-1	-1
2.	1	-1	-1	-1	-1	-1
3.	-1	1	-1	-1	-1	-1
4.	1	1	-1	-1	-1	-1
5.	-1	-1	1	-1	-1	-1
6.	1	-1	1	-1	-1	-1
7.	-1	1	1	-1	-1	-1
8.	1	1	1	-1	-1	-1
9.	-1	-1	-1	1	-1	-1
10.	1	-1	-1	1	-1	-1
11.	-1	1	-1	1	-1	-1
12.	1	1	-1	1	-1	-1
13.	-1	-1	1	1	-1	-1
14.	1	-1	1	1	-1	-1
15.	-1	1	1	1	-1	-1
16.	1	1	1	1	-1	-1
17.	-1	-1	-1	-1	1	-1
18.	1	-1	-1	-1	1	-1
19.	-1	1	-1	-1	1	-1

20.	1	1	-1	-1	1	-1
21.	-1	-1	1	-1	1	-1
22.	1	-1	1	-1	1	-1
23.	-1	1	1	-1	1	-1
24.	1	1	1	-1	1	-1
25.	-1	-1	-1	1	1	-1
26.	1	-1	-1	1	1	-1
27.	-1	1	-1	1	1	-1
28.	1	1	-1	1	1	-1
29.	-1	-1	1	1	1	-1
30.	1	-1	1	1	1	-1
31.	-1	1	1	1	1	-1
32.	1	1	1	1	1	-1
33.	-1	-1	-1	-1	-1	1
34.	1	-1	-1	-1	-1	1
35.	-1	1	-1	-1	-1	1
36.	1	1	-1	-1	-1	1
37.	-1	-1	1	-1	-1	1
38.	1	-1	1	-1	-1	1
39.	-1	1	1	-1	-1	1
40.	1	1	1	-1	-1	1
41.	-1	-1	-1	1	-1	1
42.	1	-1	-1	1	-1	1
43.	-1	1	-1	1	-1	1
44.	1	1	-1	1	-1	1
45.	-1	-1	1	1	-1	1
46.	1	-1	1	1	-1	1
47.	-1	1	1	1	-1	1
48.	1	1	1	1	-1	1
49.	-1	-1	-1	-1	1	1
50.	1	-1	-1	-1	1	1
51.	-1	1	-1	-1	1	1

52.	1	1	-1	-1	1	1
53.	-1	-1	1	-1	1	1
54.	1	-1	1	-1	1	1
55.	-1	1	1	-1	1	1
56.	1	1	1	-1	1	1
57.	-1	-1	-1	1	1	1
58.	1	-1	-1	1	1	1
59.	-1	1	-1	1	1	1
60.	1	1	-1	1	1	1
61.	-1	-1	1	1	1	1
62.	1	-1	1	1	1	1
63.	-1	1	1	1	1	1
64.	1	1	1	1	1	1

TABLE 5.2 shows the two-level full factorial matrix with all the 64 runs of all possible combinations and the matrix applied to all the data for the different years.

The tables of the two-level full factorial models for all ten years are given in appendix B.

5.3 MODEL ANALYSIS

The aim of this section is to present the results of the two-level full factorial models that were constructed over the 10 years, and to determine which demographic characteristics of the pregnant women had an effect on the risk of HIV over time.

These responses were subjected to the analysis of variance (ANOVA) procedure to investigate the effect of the various demographic characteristics and their interactions on the risk of HIV, which allowed possible variations between the main effects and the interactions between the main effects.

All possible combinations were considered and the models below show the factors and interactions that were found to have a significant effect on the risk of HIV. Some of the factors were found not to be statistically significant, but the interactions between them were found to be statistically significant, therefore the factors were included so as to ensure the correct model hierarchy.

The total degrees of freedom are equal to the total number of runs minus one for the overall mean, i.e. 64 – 1 =63. The degrees of freedom for each of the factors equal to 1 because they are considered as continuous variables.

The sum of squares of the variables represents the variability in the data that is accounted for by each variable. Therefore the total variability is indicated by the sum of squares of all the data, which is divided up into the individual sum of squares for the variables and the random error.

The F-value is the mean square for the variables in the model divided by the mean square error. The F-value indicates how much is the variability accounted for by the variable greater than the random variable.

This section also seeks to answer the second research question, namely whether the models remain stationary or change over time.

The models are represented using letters of the alphabet to represent the following:

1. M = Mother's age(pregnant woman's age)
2. F = Pregnant woman's partner's age
3. G = Gravidity
4. P = Parity
5. E = Education
6. S = Syphilis

In this section the null hypothesis being tested is:

$$Ho: B1 = B2 = B3 = B4 = B5 = B6 = 0$$

Against the alternative hypothesis

$$Ha = Bj \neq 0 \; for \; at \; least \; one \; j = 1,2,\ldots,p.$$

5.3.1 Model 2001

The section below shows the analysis of the results yielded in 2001. Tables are included.

Table 5-3: Main and interaction factors analysis, 2001

Variables	df	Parameter estimate	p-value
Intercept	1	0.485074	<.0001
Father's age	1	- 0.081209	0.0147
Education	1	-0.153986	<.0001
Gravidity	1	0.02178	0.4290
Parity	1	0.008518	0.7976
Gravidity*parity	1	-0.105514	0.0018

TABLE 5.3 shows the main effects and the interaction terms which were found to have a significant effect on pregnant women acquiring HIV.

In 2001 the age of the pregnant woman's partner, the pregnant woman's level of education and the interaction between gravidity and parity had an effect on the risk of the pregnant woman acquiring HIV.

The results above show that the pregnant woman's parity and gravidity individually did not have a significant effect on the risk of HIV.

The two-level full factorial model for 2001 is constructed as follows:

$$Y2001 = 0.485074 - 0.081209F - 0.153986E + 0.02178G + 0.008518P - 0.105514GP$$

This model shows that pregnant women who had partners of ages 26 years and older, had primary to no education and had been pregnant one or more times with no children were at a higher risk of HIV.

Table 5-4: Model statistics

Source	df	SS	MS	F	p-value
Model	5	1.628651	0.32573	7.815952	<.0001
Error	37	1.541977	0.041675		
(Lack of fit)	9	0.709589	0.078843	2.652143	0.0232
(Pure error)	28	0.832388	0.0029728		
Total	42	30170627			

Each effect shown in TABLE 5.4 is based on two averages and therefore contributes 1 degree of freedom (df) to the sum of the square, hence the $5\ df$ in the model pool. At a significance level of 0.05 the model is significant.

The F statistic has an $F(5,37)$ distribution, and according to the distribution, the chance of obtaining an F statistic of 7.816 or larger is <.0001. I therefore conclude that at least one of the six demographic characteristics of the pregnant woman is different from 0.

Table 5-5: Fit statistics, 2001

Mean	0.419708
R-square	51.37%
Adjusted R-square	44.79%

TABLE 5.5 shows that the mean for the model is 0.419708 and according to the R-square the model for 2001 explains 51.37% of the variability of the response data which is the HIV risk around the mean.

5.3.2 Model 2002

The section below gives the analysis of the results yielded for 2002. The section also includes tables.

Table 5-6: Main and interaction factors analysis, 2002

Variables	df	Parameter estimate	p-value
Intercept	1	-0.68517	<.0001
Mothage	1	0.133808	0.0128
Fathage	1	-0.031574	0.5234
Gravidity	1	-0.217572	0.0010
Syphilis	1	0.410268	<.0001
Mothage*Fathage	1	-0.215187	0.001
Gravidity*Syphilis	1	-0.129933	0.0424

TABLE 5.6 shows the results of the factors and interaction terms that had an effect on the risk of pregnant women getting HIV. The two-level full factorial model for 2002 is as follows:

$$Y2002 = -0.68517 + 0.133808M - 0.031574F - 0.217572G + 0.410268S - 0.215187MF - 0.129933GS$$

The model for 2002 shows that pregnant women of ages 25 years and older, with partners of ages 28 years and older, who had been pregnant one or more times, and tested positive for syphilis, were at a higher risk.

The interaction terms show that pregnant women of ages 28 years and older with partners of 28 years and younger were at significant risk of HIV.

The interaction between gravidity and syphilis shows that pregnant women who were pregnant for the first time and tested positive for syphilis had a significant risk of HIV.

Table 5-7: Model statistics

Source	df	SS	MS	F	p-value
Model	6	7.278586	1.213098	12.44434	<.0001

Error	34	3.314384	0.097482		
(Lack of fit)	7	1.473719	0.210531	3.088201	0.0159
(Pure error)	27	1.840665	0.068173		
Total	40	10.592997			

TABLE 5.7 shows the model's lack of fit, the model's error and its p<value (<.0001).

The F statistic has an F (6, 34) distribution. According to this distribution, the chance of obtaining an F statistic of 12.44 or larger is <.0001. I therefore conclude that at least one of the six demographic factors of the pregnant woman has an effect on the risk of HIV.

Table 5-8: Fit statistics, 2002

Mean	-0.88313
R-square	68.71%
Adjusted R-square	63.19%

TABLE 5.8 shows the model's fit statistics. The mean of the model is -0.88313, and each of the effects is based on two averages, -1 and 1, so they contributed 1 df to the sum of the square, so therefore the model is 6 df. The model for 2002 explains the 68.71% of the variability of the response data around the mean.

5.3.3 Model 2003

The section below shows the analysis of the results obtained in 2003. The section also includes tables.

Table 5-9: Main and interaction factors analysis, 2003

Variables	Df	Parameter estimate	Pr> t
Intercept	1	0.64132	
Mothage	1	-0.005019	0.9170
Fathage	1	-0.223062	0.2905
Gravidity	1	-0.223062	0.0053

Parity	1	0.103647	0.1542
Mothage*Fathage	1	-0.11621	0.0229
Gravidity*Parity	1	-0.194131	0.0137

TABLE 5.9 shows the factors as well as the interaction terms that had an effect on the risk of HIV to pregnant women for 2003. The two-level full factorial model for 2003 is as follows:

$$Y2003 = 0.64132 - 0.005019M - 0.223062F - 0.223062G + 0.103647P - 0.11621MF - 0.194131GP$$

The model for 2003 shows that mother's age and the partner's age individually had no significant impact on the risk of HIV, but the interaction between them did.

The interaction between the mother's age and the partner's age shows that pregnant women of ages 25 and younger who had partners of ages 28 years and older were at a higher risk of HIV.

The model for 2003 also showed that pregnant women experiencing their first pregnancy were at a higher risk of HIV, and that parity on its own had no significant effect on the risk of HIV, but the interaction between the pregnant woman's gravidity and parity had a significant effect on the risk of HIV.

The interaction between parity and gravidity shows that women who had been pregnant one or more times with no child were at a higher risk of HIV.

Table 5-10: Model statistics, 2003

Source	**df**	**SS**	**MS**	**F**	**p-value**
Model	6	0.954879	0.159146	2.957925	0.0327
Error	19	1.022265	0.053803		
(Lack of fit)	6	0.064864	0.010811	0.146793	0.9865
(Pure error)	13	0.9574	0.073646		
Total	25	1.977144			

TABLE 5.10 shows the model statistics for the two-level full factorial model for 2003.

The F statistic has an F (6,19) distribution and according to this distribution the chance of obtaining an F statistic of 2.96 or higher is 0.032. Therefore at least one of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

Table 5-11: Fit statistics, 2003

Mean	0.475994
R-square	48.30%
Adjusted R-square	31.97%

TABLE 5.11 shows that the model for 2003 explained 48.30% of the variation of the response data around its mean of 0.475994.

5.3.4 Model 2004

The section below shows the analysis of the results obtained for 2004. The section also includes tables.

Table 5-12: Main and interaction factors analysis, 2004

Variables	**Df**	**Parameter estimate**	**Pr> t**
Intercept	1	0.551837	<.0001
Gravidity	1	0.0486	0.1704
Parity	1	0.00713	0.8396
Education	1	-0.101713	0.0081
Gravidity*Parity	1	-0.141988	0.0002

TABLE 5.12 shows the factors and interaction terms that had a significant effect on the risk of HIV.

The two-level full factorial model for 2004 is as follows:

$$Y2004 = 0.551837 + 0.0486G + 0.00713P - 0.101713E - 0.141988GP$$

The model shows that both gravidity and parity individually had no significant effect on the risk of HIV, but the interaction between them did.

The model for 2004 showed that pregnant women with primary to no education were at a higher risk of HIV.

The interaction been gravidity and parity showed that pregnant women who had been pregnant one or more times but has no children were at a higher risk of HIV.

Table 5-13: Model statistics, 2004

Source	df	SS	MS	F	p-value
Model	4	1.407653	0.351913	6.421195	0.0004
Error	41	2.247003	0.054805		
(Lack of fit)	3	0.411243	0.137081	2.837561	0.0508
(Pure error)	38	1.83576	0.048309		
Total	45	3.654656			

TABLE 5.13 shows the model statistics for 2005 with a df of 4 and a p value<0.0004, which shows that the model is statistically significant.

The F statistic has an F (4, 41) distribution where the chance of obtaining an F statistic of 6.421 or larger is 0.0004. Therefore at least one of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

Table 5-14: Fit statistics, 2004

Mean	0.511268
R-square	38.52%
Adjusted R-square	32.52%

TABLE 5.14 shows that the model for 2004 has an R-square of 32.52%, which means that the model only accounts for 32% of the variation with a mean of 0.511268.

5.3.5 Model 2005

The section below gives the analysis of the results obtained in 2005. The section also includes tables.

Table 5-15: Main and interaction factors analysis, 2005

Variables	Df	Parameter estimate	Pr> t
Intercept	1	-0.5498	<.0001
Mothage	1	0.158406	0.0011
Fathage	1	-0.088947	0.0865
Gravidity	1	-0.203957	0.0005
Parity	1	-0.015267	0.7943
Education	1	-0.304438	<.0001
Syphilis	1	0.131806	0.0108
Mothage*Fathage	1	-0.096157	0.0320
Fathage*Education	1	0.150249	0.0067
Gravidity*Parity	1	-0.187588	0.0005
Gravidity*Syphilis	1	-0.158374	0.0031
Parity*Education	1	0.253787	<.0001
Parity*Syphilis	1	0.162595	0.0025

TABLE 5.15 shows the results for 2005 of the main effects and the interaction effects that had a significant effect on the risk of HIV.

The two-level full factorial model for 2005 is as follows:

$$Y2005 = -0.5498 + 0.158406M - 0.088947F - 0.203957G - 0.015267P - 0.304438E + 0.131806S - 0.096157MF + 0.150249FE - 0.187588GP - 0.158374GS + 0.253787PE + 0.162595PS$$

The two-level full factorial model for 2005 shows that pregnant women of ages 25 years and younger were at a higher risk of HIV.

It also showed that the pregnant woman's partner's age individually had no significant effect on the risk of HIV.

The model also shows that pregnant women who were experiencing their first pregnancy were at a higher risk of HIV, and that parity individually had no significant effect on the risk of HIV.

Pregnant women with primary to no education were at a higher risk of HIV, as well pregnant women who tested positive for syphilis.

The interaction between the woman's age and the partner's age shows that pregnant women of ages 25 years and older with partners of ages 28 years and younger were at a higher risk of HIV.

The interaction between the partner's age and the level of education of the pregnant woman shows that pregnant women with partners of ages 28 years and older with high school to tertiary education were at a higher risk of HIV.

The interaction between gravidity and parity shows that pregnant women who had one or more pregnancies, but had no children, were at a higher risk of HIV.

The interaction between gravidity and the pregnant woman's syphilis status showed that pregnant women who had one or more pregnancies before, and tested negative for syphilis, were also at a higher risk of HIV.

The interaction term between parity and the pregnant woman's level of education showed that pregnant women who had one or more children as well as secondary to tertiary education were at a higher risk of HIV, and lastly the interaction between parity and the pregnant woman's syphilis status showed that pregnant women who had one or more children and tested positive for syphilis were also at a higher risk of HIV.

Table 5-16: Model statistics, 2005

Source	df	SS	MS	F	p-value
Model	12	7.591114	0.632593	10.24097	<.0001
Error	24	1.482499	0.061771		
Total	36	9.073614			

TABLE 5.16 shows the model statistics for 2005 with a degree of freedom (df) of 12 and statistical significance at a p-value <0.0001.

The F statistic has an F (12, 24) distribution. According to this distribution, the chance of obtaining an F statistic of 10.241 or larger is <.0001. Therefore at least

one of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

Table 5-17: Fit statistics, 2005

Mean	-0.82908
R-square	83.66%
Adjusted R-square	75.49%

TABLE 5.17 shows the fit statistics for 2005. The two-level full factorial model has a mean of -0.82908 and an R-square of 83.66%, meaning that the model accounts for 84% of the variation in the data.

5.3.6 Model 2006

The section below shows the analysis of the results obtained in 2006. The section also includes tables.

Table 5-18: Main effects and interaction analysis

Variables	**Df**	**Parameter Estimate**	**Pr> t**
Intercept	1	0.557554	
Gravidity	1	-0.068481	0.0613
Parity	1	0.020357	0.5708
Education	1	-0.10599	0.0089
Syphilis	1	0.144137	0.0004
Gravidity*Parity	1	-0.13466	0.0007

TABLE 5.18 shows the main effects and interactions that had a significant effect on the risk of HIV to a pregnant woman for 2006. The two-level full factorial model for the year 2006 is as follows:

$$Year2006 = 0.557554 - 0.068481G + 0.020357P - 0.10599E + 0.144137S - 0.13466GP$$

The model for 2006 shows that the factors gravidity and parity individually had no significant effect on the risk of HIV, but the interaction between them did.

The interaction shows that pregnant women who had one or more pregnancies and had no child were at a higher risk of HIV.

The model also shows that pregnant women with primary to no education were at a higher risk of HIV, as were pregnant women who tested positive for syphilis.

Table 5-19: Model statistics, 2006

Source	df	SS	MS	F	p-value
Model	5	1.391503	0.278301	5.305233	0.0008
Error	39	2.045852	0.052458		
(Lack of fit)	8	0.706038	0.088255	2.041996	0.0739
(Pure error)	31	1.339814	0.04322		
Total	44	3.437355			

TABLE 5.19 shows the model statistics for 2006. The model has a df of 5, and shows a statistically significant p-value = 0.0008.

The F statistic has an F (5, 39) distribution, and according to this distribution the chance of obtaining an F statistic of 5.305 or larger is 0.0008. Therefore at least one of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

Table 5-20: Fit statistics, 2006

Mean	0.463022
R-square	40.48%
Adjusted R-square	32.85%

TABLE 5.20 shows that the model has a mean of 0.463022 and an R-square of 40.48%, meaning that the model explains 40% of variability of the response data around its mean.

5.3.7 Model 2007

The section below shows the analysis of the results obtained in 2007. The section also includes tables.

Table 5-21: Main and interaction factors analysis, 2007

Variables	Df	Parameter estimate	Pr> t
Intercept	1	0.451603	<.0001
Mothage	1	0.078846	0.0004
Fathage	1	0.022848	0.2711
Gravidity	1	0.049225	0.0259
Parity	1	- 0.045244	0.0397
Education	1	- 0.072078	0.0013
Syphilis	1	0.131786	<.0001
Mothage*Fathage	1	- 0.041957	0.0474
Education*Syphilis	1	- 0.06975	0.0018

TABLE 5.21 shows the results of the main effects and the interaction terms for 2007. The two-level full factorial model for 2007 is as follows:

$$Y2007 = 0.451603 + 0.078846M + 0.022848F + 0.049225G - 0.045244P - 0.072078E + 0.131786S - 0.041957MF - 0.06975ES$$

The model for 2007 shows that pregnant women of ages 25 years and older were at a higher risk of HIV, and that the pregnant woman's partner's age individually had no significant effect on the risk of the pregnant woman getting HIV, but the interaction between the mother's age and the partner's age had an effect on the risk of HIV.

The interaction between the pregnant woman's age and the partner's age showed that pregnant women of ages 25 years and older who had partners of ages 28 years and younger were at a higher risk of HIV.

The model also shows that pregnant women who had one or more pregnancies but had no children were at a higher risk of HIV, as were pregnant women who tested positive for syphilis.

The interaction between the pregnant woman's level of education and her syphilis status showed that pregnant women with primary to no education who tested positive for syphilis were at a higher risk of HIV.

Table 5-22: Model statistics, 2007

Source	df	SS	MS	F	p-value
Model	8	1.79168	0.22396	10.62919	<.0001
Error	42	0.884952	0.02107		
Total	50	2.676632			

TABLE 5.22 shows the model statistics for 2007. The model has a df of 8 and is statistically significant at a p-value<.0001.

The F statistic has an F (8, 42) distribution. According to this distribution, the chance of obtaining an F statistic of 10.629 is <.0001. Therefore at least one of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

Table 5-23: Fit statistics, 2007

Mean	0.43483
R-square	66.94%
Adjusted R-square	60.64%

TABLE 5.23 shows the fit statistics for 2007, with an R-square of 66.94%, which means that the model explains 67% of the variability of the response around its mean of 0.43483.

5.3.8 Model 2008

The section below shows the analysis of the results obtained in 2008. The section also includes tables.

Table 5-24: Main and interaction factors analysis, 2008

Variables	Df	Parameter estimate	Pr> t
Intercept	1	1.63772	<.0001
Mothage	1	- 0.147647	<.0001
Fathage	1	- 0.086745	0.0029
Gravidity	1	- 0.124135	<.0001
Parity	1	0.065893	0.0251
Education	1	0.072371	0.0125
Syphilis	1	- 0.218927	<.0001
Moth*Fathage	1	0.080371	0.0052
Fathage*Gravidity	1	0.058288	0.0382
Fathage*Syphilis	1	0.090118	0.0020
Education*Syphilis	1	0.086726	0.0033

TABLE 5.24 shows the results of the main effects and interaction terms of the model for 2008. The two-level full factorial model for 2008 is as follows:

$$Y2008 = 1.63772 - 0.147647M - 0.086745F - 0.124135G + 0.065893P + 0.072371E - 0.218927S + 0.080371MF + 0.058288FG + 0.090118FS + 0.086726ES$$

The model for 2008 shows that pregnant women of ages 25 years and younger where at a higher risk of HIV, as were pregnant women with partners of ages 28 years and younger.

The model also shows that pregnant women who were experiencing their first pregnancy and women who already had one or more children were at a higher risk of HIV. In 2008 pregnant women who had high school to tertiary education were at a higher risk of HIV.

The model also shows that even though some women tested negative for syphilis, they were still at risk of HIV due to other factors.

The model for 2008 also had interactions which tell a story. The interaction between the mother's age and partner's age shows that pregnant women of ages 25 years and above whom had partners of 28 years and older had an increased risk of HIV.

The interaction between the pregnant woman's partner's age and the pregnant woman's gravidity communicate that pregnant women who had partners of ages 28 years and older, and who had one or more pregnancies before, had an increased risk of HIV.

The interaction between the pregnant woman's partner's age and the pregnant woman's syphilis status showed that pregnant women who had partners of ages 28 years and older and tested positive for syphilis were at a higher risk of HIV.

Lastly, the interaction between the pregnant woman's level of education and her syphilis status tell us that pregnant woman who had primary to no education and who tested positive for syphilis were at a higher risk of HIV.

Table 5-25: Model statistics, 2008

Source	df	SS	MS	F	p-value
Model	10	5.748405	0.574841	15.57255	<.0001
Error	40	1.476548	0.036914		
Total	50	7.224953			

TABLE 5.25 shows model statistics for 2008 with a df of 10 and statistically significant with a p-value <.0001.

The F statistic has an F (10, 40) distribution, and according to this distribution the chance of getting an F statistic of 15.572 or larger is <.0001. Therefore at least one of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

Table 5-26: Fit statistics, 2008

Mean	1.653142
R-square	79.56%
Adjusted R-square	74.45%

TABLE 5.26 shows the fit statistics for 2008 with an R-square that explains or accounts for 79.56% of the variability of the response data around a mean of 1.653142.

5.3.9 Model 2009

The section below shows the analysis of the results obtained in 2009. The section also includes tables.

Table 5-27: Main and interaction factors analysis, 2009

Variables	**Df**	**Parameter Estimate**	**Pr> t**
Intercept	1	0.732649	
Mothage	1	0.02691	0.2062
Fathage	1	-0.037265	0.0869
Gravidity	1	-0.041184	0.0573
Parity	1	0.026329	0.2090
Education	1	-0.085913	0.0003
Syphilis	1	0.111103	<.0001
Mothage*Fathage	1	-0.057569	0.0095
Fathage*Syphilis	1	-0.061598	0.0060
Gravidity*Parity	1	-0.081024	0.0003
Gravidity*Education	1	0.057734	0.0090

TABLE 5.27 shows the main effects as well as the interaction for the two-level full factorial model for 2008. The full factorial model for the year 2008 is as follows:

$$Y2009 = 0.732649 + 0.02691M - 0.037265F - 0.041184G + 0.026329P - 0.085913E + 0.111103S - 0.057569MF - 0.061598FS - 0.081024GP + 0.057734GE$$

The two-level full factorial model for 2009 shows that the mother's age, the pregnant woman's partner's age and the pregnant woman's gravidity and parity individually had no significant effect on the risk of HIV, but the interactions between these variables did.

The model also shows that pregnant women with primary to no education had a higher risk of HIV, as did pregnant women who tested positive for syphilis.

The interaction between the mother's age and the partner's age shows that the risk of HIV increased in pregnant women of ages 25 years and older with partners of ages 28 years and younger.

The interaction between the pregnant woman's partner's age and her syphilis status showed that pregnant women who had partners of ages 28 years and older and tested negative for syphilis had an increased risk of HIV.

The interaction between the pregnant woman's gravidity and parity showed that those who had one or more pregnancies previously but had no children were at a higher risk of HIV.

Lastly, the interaction between the pregnant woman's gravidity and level of education showed that pregnant women who had had one or more pregnancies previously and had primary to no education were at a higher risk of HIV.

Table 5-28: Model statistics, 2009

Source	df	SS	MS	F	p-value
Model	10	1.546674	0.154667	7.312031	<.0001
Error	40	0.846098	0.021152		
Total	50	2.392771			

TABLE 5.28 shows the model statistics for 2009, with a df of 10 and a statistically significant p-value < .0001.

The F statistic has an F (10, 40) distribution. According to this distribution the chance of obtaining an F statistic of 7.312 is <.0001.

Therefore at least one of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

Table 5-29: Fit statistics, 2009

Mean	0.682155
R-square	64.64%
Adjusted R-square	55.80%

TABLE 5.29 represents the fit statistics for 2009 which have an R-square of 64.64%, meaning that it accounts for 64% of the variability of the response data around a mean of 0.682155.

5.3.10 Model 2010

The section below shows the analysis of the results obtained in 2010. The section also includes tables.

Table 5-30: Main and interaction factors analysis, 2010

Variables	**Df**	**Parameter estimate**	**Pr> t**
Intercept	1	-0.68935	<.0001
Mothage	1	0.147483	0.0333
Fathage	1	-0.081799	0.2440
Gravidity	1	-0.017999	0.7899
Parity	1	0.086465	0.2025
Education	1	-0.203096	0.0077
Syphilis	1	0.304817	0.0001
Mothage*Fathage	1	-0.156449	0.0226

Fathage*Education	1	0.189976	0.0084
Gravidity*Parity	1	-0.208332	0.0040

TABLE 5.30 shows the results of the two-level full factorial model for 2010. It shows the main effects as well as the interactions between the variables. The model for the year 2010 is as follows:

$$Y2010 = -0.68935 + 0.147483M - 0.081799F - 0.017999G + 0.086465P - 0.203096E + 0.304817S - 0.156449MF + 0.189976FE - 0.208332GP$$

The two-level full factorial model for 2010 show that the factors pregnant woman's partner's age, gravidity and parity individually have no significant impact on the risk of HIV, but the interactions between the variables do.

The model for 2010 shows that pregnant women of ages 25 years and older were at a higher risk of HIV, and pregnant women who had primary to no education were at a higher risk of HIV, as were pregnant women who tested positive for syphilis.

The interaction between the mother's age and the pregnant woman's partner's age showed that pregnant women of ages 25 years and older who had partners of age 28 years and younger were at a higher risk of HIV.

The interaction between the pregnant woman's partner's age and the pregnant woman's level of education showed that pregnant women who had partners of ages 28 years and older, with secondary to tertiary education were at a higher risk of HIV.

The interaction between the pregnant woman's gravidity and parity showed that pregnant women who had had one or more pregnancies previously but had no children were at a higher risk of HIV.

Table 5-31: Model statistics, 2010

Source	**df**	**SS**	**MS**	**F**	**p-value**
Model	9	8.470886	0.94121	4.736476	0.0003
Error	37	7.352461	0.198715		
Total	46	15.82335			

TABLE 5.31 shows the model statistics for 2010 with a df of 9 and a statistically significant p-value of 0.0003.

The F statistic has an F (9, 37) distribution. According to this distribution, the chance of obtaining an F statistic of 4.736 or larger is 0.0003.

Therefore at least one of the demographic characteristics of the pregnant woman has an effect on the risk of HIV.

Table 5-32: Fit statistics, 2010

Mean	-0.86899
R-square	53.53%
RMSE	0.445775
CV	-51.2982

TABLE 5.32 shows the fit statistics for 2010 with an R-square which accounts for 53.53% of the variability of the data around a mean of -0.86899.

5.3.11 Overall analysis of models

The second research question asked whether the model remains stationary or not.

The aim was to analyse the models and determine whether the factors and their interactions which had an effect on the risk of HIV changed over time or remained the same.

Table 5-33: Two-level full factorial models, 2001 to 2010

Two-level full factorial models	Model Output
Model 2001	Y2001 = 0.485074 - 0.081209F - 0.153986E + 0.02178G +0.008518P - 0.105514GP

Model 2002	Y2002 = -0.68517 + 0.133808M - 0.031574F -0.217572G + 0.410268S - 0.215187MF -0.129933GS
Model 2003	Y2003 = 0.64132 -0.005019M - 0.223062F -0.223062G + 0.103647P - 0.11621 MF -0.194131GP
Model 2004	Y2004 = 0.551837 + 0.0486G + 0.00713P -0.101713E -0.141988GP
Model 2005	Y2005 = -0.5498 + 0.158406M - 0.088947F -0.203957G -0.015267P - 0.304438E + 0.131806S -0.096157MF + 0.150249FE -0.187588GP -0.158374GS + 0.253787PE + 0.162595PS
Model 2006	Year2006 = 0.557554 - 0.068481G + 0.020357P - 0.10599E + 0.144137S - 0.13466GP
Model 2007	Y2007= 0.451603 + 0.078846M + 0.022848F + 0.049225G - 0.045244P - 0.072078E + 0.131786S - 0.041957MF - 0.06975ES
Model 2008	Y2008 = 1.63772 - 0.147647M - 0.086745F - 0.124135G + 0.065893P + 0.072371E - 0.218927S + 0.080371MF + 0.058288FG + 0.090118FS + 0.086726ES
Model 2009	Y2009= 0.732649 + 0.02691M - 0.037265F - 0.041184G + 0.026329P - 0.085913E + 0.111103S - 0.057569MF - 0.061598FS - 0.081024GP + 0.057734GE
Model 2010	Y2010 = - 0.68935 + 0.147483M - 0.081799F - 0.017999G + 0.086465P - 0.203096E + 0.304817S - 0.156449MF + 0.189976FE - 0.208332GP

TABLE 5.33 shows all 10 of the full factorial models. It shows that the risk of HIV of the pregnant women was affected by different factors and interactions from year to year.

The model for 2001 is not the same as the one for 2002. This could be due to the number of people who were tested each year or the different locations from which the data was collected, but the model does not remain stationary from year to year.

5.4 CONCLUSION

The objective of this chapter was to demonstrate an understanding of the actual development of the two-level full factorial model, as well as to provide the two-level full factorial models for every year. This objective was achieved – a brief description was provided of what a two-level full factorial model is, as well as the actual development of a two-level full factorial model for each year, which is given in Appendix B. The objective was also achieved by using ANOVA to analyse the HIV data results produced, which also assisted in the development of the HIV models for the various years. This chapter was also able to answer the research question as to whether the model remains stationary or changes over the years by providing a summary of the models and an analysis of whether the model had changed or remained the same.

The next chapter provides the conclusions of the study and recommendations.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

6.1 INTRODUCTION

The primary objective of this study was to develop a two-level full factorial model to analyse HIV data. In order to achieve this, the study formulated a Design of Experiments (DOE) methodology discussed in CHAPTER 2. This chapter discusses the final proceedings of this study which are the conclusions and recommendations.

The objective of this chapter is to communicate the findings of this study and bring it to a conclusion. The chapter summarises all the key concepts from the previous chapters, and answers the research questions posed by this study. The objectives of the study are addressed.

This chapter is divided into the following sections: summary of the research findings of the study (Section 6.2); recommendations for future research (Section 6.3); and finally the closure of the study (Section 5.4).

6.2 RESEARCH FINDINGS OF THE STUDY

This section revisits the research question, the primary objective and the theoretical objectives and provides information on how these were addressed in the study.

The research questions for this study are:

- Is it possible to develop a two-level full factorial model to analyse HIV data?
- Does the model remain stationary or does it change over time?

These research questions are supported by the following primary objective:

- The primary objective of this study is to develop two-level full factorial models with which to analyse antenatal HIV data on an annual basis.

The primary objective is supported by the following theoretical objectives:

- Search the literature to gain a better understanding of the Design of Experiments methodology.
- Search the literature to gain a better understanding of two-level full factorial models
- Search the literature to gain a better understanding of data analysis

The following sections reflect on the research questions, the primary objective and the theoretical objectives by highlighting key findings.

Theoretical objectives

The sections below represent key findings based on the literature review during the study.

5.2.1 Design of Experiments

The development of a two-level full factorial model is an important requirement of the primary objective. The theoretical objective of DOE was discussed through a review of existing literature in CHAPTER 2: addressing the first theoretical objective.

The study found that DOE was first developed by Ronald A. Fisher in his Rothmans laboratory for agricultural purposes, but the methodology has since then been applied in various industries. The study also found that DOE is structured in an orderly way to conduct experiments as well as to analyse how the factors in question affect the outcome of the response variable.

Understanding the value of DOE is important to this study, therefore the philosophical position taken is that of positivism with DOE as a methodology.

DOE provides the researcher with the flexibility of comparing more than one factor at a time, and also provides the opportunity to consider all possible combinations (Telford, 2007).

5.2.2 Statistical Methods

The objective of CHAPTER 3 was to give a literature review on the statistical methods applied in the study. This was achieved by providing a brief history of the data was given in SECTION 3.2. The data was collected by the National Department of Health (NDoH) with the use of their yearly survey which is conducted on all pregnant women attending an antenatal clinic for the first time in all nine provinces in all 52 health districts. The survey is used as a tool to assist the government to determine the prevalence of HIV among pregnant women in South Africa.

A brief summary of similar studies conducted was included, which found that the Sub-Saharan region in Africa is the most affected by the HIV epidemic. A study conducted by Lawi et al. (2015a) found that it is necessary to re-screen women after they have given birth so as to avoid missed opportunities of identifying syphilis and HIV that may have been contracted during pregnancy. SECTION 3.3 took a closer look at contingency table, which assists in grouping the data orderly. The chi-square test is used in contingency tables to compare the observed count in each table cell to the number which is expected under the assumption of no association between the rows and column classification (Diener-West, 2008). These methods were applied in CHAPTER 4 to better understand and interpret the pregnant woman's demographic characteristics.

SECTION 3.4 gave an overview on time series which was defined by Swanepoel et al. (2011) as observations which are collected over time.

Time series assist us to monitor how certain variables change over the course of time. As already mentioned, this study attempted to determine change to the risk of pregnant women acquiring HIV over time by analysing their demographic characteristics over a ten-year period. Linear models were applied in CHAPTER 4 to analyse the pregnant woman's demographical characteristic, and the coefficient of determination was used to determine how much variability the models account for.

SECTION 3.5 took a closer look at Anova, which is defined as a multivariate method used to analyse variation in a response variable normally used to test equality among means by comparing variance among groups relative to variance within groups (Larson, 2008). This section also defined the various measures used in Anova, which are the sum of squares, as well as defined the assumptions found in Anova, namely independence, normal distribution and variation.

5.2.2 Primary objective: Data analysis and interpretation

The sections that follow discuss key findings concerning the development of two-level full factorial models to analyse HIV antenatal data over time, which addresses the primary objective.

5.2.3 Data analysis

CHAPTER 4 analyses the pregnant woman's demographic characteristics with the use of linear models. The study found the use of linear models was useful in understanding the demographic characteristics of pregnant women, as the comparisons used gave a broader view of the factors.

Chapter 4 took a closer look at all the demographic characteristics of the pregnant woman with the aim of understanding the data and its variables. It was found that:

Pregnant women of ages 25 years and older were at a higher risk of HIV than those of ages 25 years and younger. The study also found that both the mother's age and the father's age experienced an upward trend over the ten-year period.

The results also found that pregnant women with partners of ages 28 years and older were at a higher risk of HIV, as compared to those with partners of ages 28 years and younger.

Pregnant women who had previously had one or more pregnancies were found to be at higher risk of HIV than those who were experiencing their first pregnancy.

Pregnant women who had one or more children were also found to be at higher risk than pregnant women who had no children at all.

The level of education of the pregnant woman experienced an upward trend in both pregnant women who had primary to no education and those who had secondary to tertiary education. Although this is the case, pregnant women with primary to no education were at a higher risk of HIV.

Pregnant women who tested positive for syphilis were at a higher risk of HIV than pregnant women who tested negative for syphilis.

5.2.4 Development of two-level full factorial models

CHAPTER 5 looked at the development of the two-level full factorial model and whether it remained stationary over the ten-year period: addressing the second research question.

SECTION 5.2 described a factorial design as a structure where the effects of many different factors or treatments are tested at the same time (Morris, 2011). In this study all the demographic characteristics of the pregnant woman were tested and processed simultaneously.

TABLE 5.2 shows the design of a two-level full factorial design matrix. The tables of the two-level full factorial models that were generated for the period 2001 to 2010 are included in APPENDIX B.

TABLE 5.33 gives all the HIV models from 2001 to 2010, and it was concluded that the models changed from year to year; therefore the two-level full factorial models are not stationary but change according to which factors were most prevalent in that year.

CHAPTER 5 answers the question whether the model remains stationary or not, and the results showed that the model changed over time. The risk of HIV is not affected by just one variable, but multiple variables and their interactions, therefore the models change every year.

6.3 RECOMMENDATIONS FOR FUTURE RESEARCH

Various research possibilities were identified during the course of this study. These possibilities are improvements that may be made to the study, and the possibility of expanding the study for PhD purposes. The possibilities include the following:

a. The study is limited to coded data, and attaining the full data could improve the results of the analysis of the data.
b. The study looked at a wide range of data, but narrowing down the number of years and the number of variables may yield interesting results.
c. Future research could focus on other Design of Experiment methods, such as fractional design.
d. More recent data for analysis could be acquired, and forecasting could be done.
e. Restrictions to the data for future studies could be imposed.
f. Future research could take a closer look at just the interactions between the factors.

6.4 CLOSURE OF THE STUDY

The aim of this study was to develop a two-level full factorial model for the analysis of HIV antenatal data. This was achieved by reviewing the existing literature on DOE, as well as taking a closer look at the pregnant woman's demographic characteristics in CHAPTER 4. This was also achieved by the actual development of two-level full factorial models for each year in CHAPTER 5.

The study found that the use of linear models, line charts and tables in CHAPTER 4 was particularly helpful in understanding the demographic characteristics of the pregnant women, as well as being able to determine the trend of the prevalence of HIV among the pregnant women.

The development of the full factorial models was of extreme importance in bringing the entire study together. ANOVA analysis was particularly helpful in developing the models for the ten years. The development of a two-level full factorial model for each year assisted in determining the risk of HIV over time.

The study concludes that it is possible to develop two-level full factorial models for the analysis of HIV antenatal data over time, and that the model does not remain stationary and changes from year to year.

The model re-emphasised that mother's age, father's age, parity, gravidity and syphilis are the most common combinations that played a major role in the risk of HIV. The study was also able to fill a gap in the literature which was to develop a two-level full factorial model using HIV antenatal data. The study was also able to give a broader view and understanding factors that affect the risk of HIV over time.

CHAPTER 7: REFERENCES

ALIYU, A. A., BELLO, M. U., KASIM, R. & MARTIN, D. 2014. Positivist and Non-Positivist Paradigm in Social Science Research: Conflicting Paradigms or Perfect Partners? *Journal of Management and Sustainability,* 4**,** 79-95.

ALLEN, M. P. 2007. *Understanding regression analysis*, Springer Science & Business Media.

ANDERSON, M. J. & WHITCOMB, P. J. 2015. *DOE simplified: practical tools for effective experimentation*, CRC Press.

BINGÖL, D., SARAYDIN, D. & ÖZBAY, D. Ş. 2015. Full Factorial Design Approach to Hg (II) Adsorption onto Hydrogels. *Arabian Journal for Science and Engineering,* 40**,** 109-116.

BOON, M. & MARIATTI, M. 2014. Optimization of magnetic and dielectric properties of surface-treated magnetite-filled epoxy composites by factorial design. *Journal of Magnetism and Magnetic Materials,* 355**,** 319-324.

CAMPBELL, D. & CAMPBELL, S. Introduction to regression and data analysis. StatLap Workshop Series, 2008. 1-15.

CAVAZZUTI, M. 2013. Design of experiments. *Optimization Methods.* Springer.

CHALEUNVONG, K. 2009. Data collection techniques. *Training Course in Reproductive Health Research Vientine*.

CRESWELL, J. W. 2013. A Framework for Design. *In:* AXELSEN, D. E. (ed.) *Research design: Qualitative, quantitative, and mixed methods approaches*

London: Sage publications.

GOGTAY, N., DESHPANDE, S. & THATTE, U. 2017. Principles of Regression Analysis. *Journal of The Association of Physicians of India,* 65**,** 48.

GOR, R. M. 2009. 6: Forecasting Techniques.

GUBA, E. G. & LINCOLN, Y. S. 1994. Competing paradigms in qualitative research. *Handbook of qualitative research***,** 105-117.

GUPTA, V. & PARSAD, R. 2006. Fundamentals of Design of Experiments. *Statistical Manual,* 2**,** 199-162.

HEVNER, R., MARCH, S., PARK, J. & RAM, S. 2004. Design science in information systems research. *MIS quarterly,* 28**,** 75-105.

ISOTALO, J. 2001. Basics of statistics. *Finland: University of Tampere*.

JAYNES, J. L. 2013. *Contributions in Design of Experiments: Methods and Applications.* Doctor of Philosophy in Statistics Thesis -PhD, University of California Los Angeles.

JMP 2014. *Design of Experiments Guide,* North Carolina, SAS Institute Inc.

KRAUSS, S. E. 2005. Research paradigms and meaning making: A primer. *The qualitative report,* 10**,** 758-770.

KULATUNGA, K., AMARATUNGA, R. & HAIGH, R. 2007. Researching construction client and innovation: methodological perspective.

LARSON, M. G. 2008. Analysis of variance. *Circulation,* 117**,** 115-121.

LAWAL, B. 2014. *Applied statistical methods in agriculture, health and life Sciences*, Springer.

LAWI, J. D., MIRAMBO, M. M., MAGOMA, M., MUSHI, M. F., JAKA, H. M., GUMODOKA, B. & MSHANA, S. E. 2015a. Sero-conversion rate of Syphilis

and HIV among pregnant women attending antenatal clinic in Tanzania: a need for re-screening at delivery. *BMC pregnancy and childbirth,* 15**,** 7.
LAWI, J. D., MIRAMBO, M. M., MAGOMA, M., MUSHI, M. F., JAKA, H. M. & GUMODOKA, B., MSHANA STEPHEN E. 2015b. Sero-conversion rate of Syphilis and HIV among pregnant women attending antenatal clinic in Tanzania: a need for re-screening at delivery. *Bugando Medical Centre(BMC) Pregnancy and Childbirth,* 15**,** 1-7.
LEWIS-BECK, M., BRYMAN, A. E. & LIAO, T. F. 2003. *The Sage encyclopedia of social science research methods*, Sage Publications.
LEWIS, P., THORNHILL, A. & SAUNDERS, M. 2007. *Research methods for business students*, Pearson Education UK.
MANIKANDAN, S. 2011. Frequency distribution. *Journal of pharmacology & pharmacotherapeutics,* 2**,** 54.
MEE, R. 2009. *A comprehensive guide to factorial two-level experimentation*, Springer Science & Business Media.
MONTGOMERY, D. C. 2017. *Design and analysis of experiments*, John Wiley & Sons.
MOORE, D. S., MCCABE, G. P. & CRAIG, B. A. 2012. Introduction to the Practice of Statistics.
MORRIS, M. 2011. Factorial treament structure. *In:* CARLIN, B. P., FARAWAY, J. J., TANNER, M. & ZIDEK, J. (eds.) *Design of experiments: An introduction based on linear models*

Florida: CRC Press.
NDOH 2012. The 2011 National Antenatal Sentinel HIV and Syphilis Prevalence Survey in South Africa. *In:* HEALTH, N. D. O. (ed.). Pretoria: National Department of Health
NOOR, K. B. 2008. Case study: A strategic research methodology. *American journal of applied sciences,* 5**,** 1602-1604.
OEHLERT, G. W. 2010. A first course in design and analysis of experiments.
PECK, R., OLSEN, C. & DEVORE, J. 2015. *Introduction to statistics and data analysis*, Cengage Learning.
PEFFERS, K., TUUNANEN, T., ROTHENBERGER, M. A. & CHATTERJEE, S. 2007. A design science research methodology for information systems research. *Journal of management information systems,* 24**,** 45-77.
RAJASEKAR, S., PHILOMINATHAN, P. & CHINNATHAMBI, V. 2013. Research Methodology.
RAWLINGS, J. O., PANTULA, S. G. & DICKEY, D. A. 2001. *Applied regression analysis: a research tool*, Springer Science & Business Media.
RAY, S. 2017. 7 Types of Regression Techniques you should know!
SAUNDERS, M., LEWIS, P. & THORNHILL, A. 2009. *Research methods for business students,* 5th ed. Harlow, Pearson
SCOTLAND, J. 2012. Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and critical research paradigms. *English Language Teaching,* 5**,** 9-16.
SELTMAN, H. J. 2012. Experimental design and analysis. *Online at: http://www. stat. cmu. edu/, hseltman/309/Book/Book. pdf*.
SIBANDA, W. 2013. *Comparative Study of Neural Networks and Design of Experiments to the Classification of HIV status.* Dr of Philosophy in Information Technology Thesis-PhD, North-West University.

SIBANDA, W. & PRETORIUS, P. 2011. Application of Two-level Fractional Factorial Design to Determine and Optimize the Effect of Demographic Characteristics on HIV Prevalence using the 2006 South African Annual Antenatal HIV and Syphilis Seroprevalence data. *International Journal of Computer Applications,* 35**,** 15-20.

SIBANDA, W. & PRETORIUS, P. 2014. Weight of Evidence Model: Application in Tracking Changes in HIV Risk Profile Using 10 Year Annual Antenatal HIV Seroprevalence Data. *International Journal of Biochemistry and Bioinformatics,* 4**,** 212-217.

STEYN, H. S. & SWANEPOEL, C. J. 2008. *Practical Statistics*.

STOKES, M. E., DAVIS, C. S. & KOCH, G. G. 2012. *Categorical data analysis using SAS*, SAS institute.

SWANEPOEL, J., SWANEPOEL, C., ALLISON, J. & SANTANA, L. 2011. *Elementary Statistical Methods 4TH EDITION*, AndCork Publishers.

TELFORD, J. K. 2007. A brief introduction to design of experiments. *Johns Hopkins apl technical digest,* 27**,** 224-232.

VAISHNAVI, V. & KUECHLER, W. 2004. *Design research in information systems* [Online]. Available: http://desrist.org/design-research-in-information-systems/ [Accessed 30 May 2014].

WAHYUNI, D. 2012. The research design maze: Understanding paradigms, cases, methods and methodologies. *Journal of Applied Management Accounting Research,* 10**,** 69-80.

YONG, A. G. & PEARCE, S. 2013. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology,* 9**,** 79-94.

APPENDIX A: FREQUENCY PROCEDURE

APPENDIX A1: Frequency procedure by mother's age (2001 -2010)

HIVclass by Mothage 2001			
HIVclass)	Mothage		
	-1	1	Total
0	4506	5002	9508
	36.78	40.83	77.62
	47.39	52.61	
	77.00	78.18	
1	1346	1396	2742
	10.99	11.40	22.38
	49.09	50.91	
	23.00	21.82	
Total	5852	6398	12250
	47.77	52.23	100.00

Table of HIVclass by Mothage 2002			
HIVclass(HIVclass)	Mothage(Mothage)		
	-1	1	Total
0	5888 38.40 52.31 76.55	5368 35.01 47.69 70.25	11256 73.41
1	1804 11.77 44.25 23.45	2273 14.82 55.75 29.75	4077 26.59
Total	7692 50.17	7641 49.83	15333 100.00

Table of HIVclass by Mothage 2003			
HIVclass(HIVclass)	Mothage(Mothage)		
	-1	1	Total
0	675 36.10 49.41 73.05	691 36.95 50.59 73.04	1366 73.05
1	249 13.32 49.40 26.95	255 13.64 50.60 26.96	504 26.95
Total	924 49.41	946 50.59	1870 100.00

Table of HIVclass by Mothage 2004			
HIVclass(HIVclass)	Mothage(Mothage)		
	-1	1	Total
0	5576 37.63 53.34 74.90	4878 32.92 46.66 66.18	10454 70.56
1	1869 12.61 42.85 25.10	2493 16.83 57.15 33.82	4362 29.44
Total	7445 50.25	7371 49.75	14816 100.00

Table of HIVclass by Mothage 2005			
HIVclass(HIVclass)	Mothage(Mothage)		
	-1	1	Total
0	5061 39.57 56.00 75.59	3976 31.09 44.00 65.23	9037 70.66
1	1634 12.78 43.54 24.41	2119 16.57 56.46 34.77	3753 29.34
Total	6695 52.35	6095 47.65	12790 100.00

Table of HIVclass by Mothage 2006			
HIVclass(HIVclass)	**Mothage(Mothage)**		
	-1	**1**	**Total**
0	12508	9761	22269
	40.02	31.23	71.24
	56.17	43.83	
	77.74	64.35	
1	3582	5407	8989
	11.46	17.30	28.76
	39.85	60.15	
	22.26	35.65	
Total	16090	15168	31258
	51.47	48.53	100.00

Table of HIVclass by Mothage 2007			
HIVclass(HIVclass)	**Mothage(Mothage)**		
	-1	**1**	**Total**
0	12947 39.86 56.15 77.92	10112 31.13 43.85 63.73	23059 70.99
1	3668 11.29 38.93 22.08	5754 17.71 61.07 36.27	9422 29.01
Total	16615 51.15	15866 48.85	32481 100.00

Table of HIVclass by Mothage 2008			
HIVclass(HIVclass)	**Mothage(Mothage)**		
	-1	**1**	**Total**
0	13081 40.16 56.63 78.37	10020 30.76 43.37 63.10	23101 70.93
1	3611 11.09 38.13 21.63	5859 17.99 61.87 36.90	9470 29.07
Total	16692 51.25	15879 48.75	32571 100.00

Table of HIVclass by Mothage 2009			
HIVclass(HIVclass)	Mothage(Mothage)		
	-1	1	Total
0	12387 39.77 56.28 78.41	9622 30.90 43.72 62.70	22009 70.67
1	3410 10.95 37.33 21.59	5724 18.38 62.67 37.30	9134 29.33
Total	15797 50.72	15346 49.28	31143 100.00

Table of HIVclass by Mothage 2010			
HIVclass(HIVclass)	Mothage(Mothage)		
	-1	1	Total
0	11942 39.66 56.63 78.46	9147 30.38 43.37 61.43	21089 70.04
1	3279 10.89 36.34 21.54	5744 19.08 63.66 38.57	9023 29.96
Total	15221 50.55	14891 49.45	30112 100.00

APPENDIX A2: Frequency procedure by father's age (2001 – 2010)

Table of HIVclass by Fathage 2001			
HIVclass(HIVclass)	Fathage(Fathage)		
	-1	1	Total
0	2435 19.88 25.61 82.21	7073 57.74 74.39 76.15	9508 77.62
1	527 4.30 19.22 17.79	2215 18.08 80.78 23.85	2742 22.38
Total	2962 24.18	9288 75.82	12250 100.00

Table of HIVclass by Fathage 2002			
HIVclass(HIVclass)	Fathage(Fathage)		
	-1	1	Total
0	5547 36.18 49.28 77.31	5709 37.23 50.72 69.98	11256 73.41
1	1628 10.62 39.93 22.69	2449 15.97 60.07 30.02	4077 26.59
Total	7175 46.79	8158 53.21	15333 100.00

Table of HIVclass by Fathage 2003			
HIVclass(HIVclass)	Fathage(Fathage)		
	-1	1	Total
0	625 33.42 45.75 74.94	741 39.63 54.25 71.53	1366 73.05
1	209 11.18 41.47 25.06	295 15.78 58.53 28.47	504 26.95
Total	834 44.60	1036 55.40	1870 100.00

Table of HIVclass by Fathage 2004			
HIVclass(HIVclass)	Fathage(Fathage)		
	-1	1	Total
0	5273 35.59 50.44 75.71	5181 34.97 49.56 65.99	10454 70.56
1	1692 11.42 38.79 24.29	2670 18.02 61.21 34.01	4362 29.44
Total	6965 47.01	7851 52.99	14816 100.00

Table of HIVclass by Fathage 2005			
HIVclass(HIVclass)	Fathage(Fathage)		
	-1	1	Total
0	4868 38.06 53.87 77.05	4169 32.60 46.13 64.42	9037 70.66
1	1450 11.34 38.64 22.95	2303 18.01 61.36 35.58	3753 29.34
Total	6318 49.40	6472 50.60	12790 100.00

Table of HIVclass by Fathage 2006			
HIVclass(HIVclass)	Fathage(Fathage)		
	-1	1	Total
0	11797 37.74 52.97 78.18	10472 33.50 47.03 64.77	22269 71.24
1	3292 10.53 36.62 21.82	5697 18.23 63.38 35.23	8989 28.76
Total	15089 48.27	16169 51.73	31258 100.00

Table of HIVclass by Fathage 2007			
HIVclass(HIVclass)	**Fathage(Fathage)**		
	-1	**1**	**Total**
0	12437 38.29 53.94 78.39	10622 32.70 46.06 63.93	23059 70.99
1	3428 10.55 36.38 21.61	5994 18.45 63.62 36.07	9422 29.01
Total	15865 48.84	16616 51.16	32481 100.00

Table of HIVclass by Fathage 2008			
HIVclass(HIVclass)	Fathage(Fathage)		
	-1	1	Total
0	12696	10405	23101
	38.98	31.95	70.93
	54.96	45.04	
	78.56	63.40	
1	3464	6006	9470
	10.64	18.44	29.07
	36.58	63.42	
	21.44	36.60	
Total	16160	16411	32571
	49.61	50.39	100.00

Table of HIVclass by Fathage 2009			
HIVclass(HIVclass)	Fathage(Fathage)		
	-1	1	Total
0	11914	10095	22009
	38.26	32.41	70.67
	54.13	45.87	
	78.69	63.08	
1	3226	5908	9134
	10.36	18.97	29.33
	35.32	64.68	
	21.31	36.92	
Total	15140	16003	31143
	48.61	51.39	100.00

Table of HIVclass by Fathage 2010			
HIVclass(HIVclass)	Fathage(Fathage)		
	-1	1	Total
0	11839	9250	21089
	39.32	30.72	70.04
	56.14	43.86	
	78.41	61.61	
1	3259	5764	9023
	10.82	19.14	29.96
	36.12	63.88	
	21.59	38.39	
Total	15098	15014	30112
	50.14	49.86	100.00

APPENDIX A3: Frequency procedure by gravidity (2001 – 2010)

Table of HIVclass by Gravidity 2001			
HIVclass(HIVclass)	Gravidity(Gravidity)		
	-1	1	Total
0	3752	5756	9508
	30.63	46.99	77.62
	39.46	60.54	
	78.22	77.23	
1	1045	1697	2742
	8.53	13.85	22.38
	38.11	61.89	
	21.78	22.77	
Total	4797	7453	12250
	39.16	60.84	100.00

Table of HIVclass by Gravidity 2002			
HIVclass(HIVclass)	Gravidity(Gravidity)		
	-1	1	Total
0	323	10933	11256
	2.11	71.30	73.41
	2.87	97.13	
	73.24	73.42	
1	118	3959	4077
	0.77	25.82	26.59
	2.89	97.11	
	26.76	26.58	
Total	441	14892	15333
	2.88	97.12	100.00

Table of HIVclass by Gravidity 2003			
HIVclass(HIVclass)	Gravidity(Gravidity)		
	-1	1	Total
0	583	783	1366
	31.18	41.87	73.05
	42.68	57.32	
	73.33	72.84	
1	212	292	504
	11.34	15.61	26.95
	42.06	57.94	
	26.67	27.16	
Total	795	1075	1870
	42.51	57.49	100.00

Table of HIVclass by Gravidity 2004			
HIVclass(HIVclass)	Gravidity(Gravidity)		
	-1	1	Total
0	4584	5870	10454
	30.94	39.62	70.56
	43.85	56.15	
	76.22	66.69	
1	1430	2932	4362
	9.65	19.79	29.44
	32.78	67.22	
	23.78	33.31	
Total	6014	8802	14816
	40.59	59.41	100.00

Table of HIVclass by Gravidity 2005			
HIVclass(HIVclass)	Gravidity(Gravidity)		
	-1	1	Total
0	4003	5034	9037
	31.30	39.36	70.66
	44.30	55.70	
	77.85	65.82	
1	1139	2614	3753
	8.91	20.44	29.34
	30.35	69.65	
	22.15	34.18	
Total	5142	7648	12790
	40.20	59.80	100.00

Table of HIVclass by Gravidity 2006			
HIVclass(HIVclass)	**Gravidity(Gravidity)**		
	-1	**1**	**Total**
0	9729 31.12 43.69 79.24	12540 40.12 56.31 66.07	22269 71.24
1	2549 8.15 28.36 20.76	6440 20.60 71.64 33.93	8989 28.76
Total	12278 39.28	18980 60.72	31258 100.00

Table of HIVclass by Gravidity 2007			
HIVclass(HIVclass)	**Gravidity(Gravidity)**		
	-1	**1**	**Total**
0	9977 30.72 43.27 79.88	13082 40.28 56.73 65.44	23059 70.99
1	2513 7.74 26.67 20.12	6909 21.27 73.33 34.56	9422 29.01
Total	12490 38.45	19991 61.55	32481 100.00

Table of HIVclass by Gravidity 2008			
HIVclass(HIVclass)	Gravidity(Gravidity)		
	-1	1	Total
0	10186	12915	23101
	31.27	39.65	70.93
	44.09	55.91	
	80.32	64.94	
1	2496	6974	9470
	7.66	21.41	29.07
	26.36	73.64	
	19.68	35.06	
Total	12682	19889	32571
	38.94	61.06	100.00

Table of HIVclass by Gravidity 2009			
HIVclass(HIVclass)	Gravidity(Gravidity)		
	-1	1	Total
0	9522	12487	22009
	30.58	40.10	70.67
	43.26	56.74	
	80.55	64.63	
1	2299	6835	9134
	7.38	21.95	29.33
	25.17	74.83	
	19.45	35.37	
Total	11821	19322	31143
	37.96	62.04	100.00

Table of HIVclass by Gravidity 2010			
HIVclass(HIVclass)	**Gravidity(Gravidity)**		
	-1	**1**	**Total**
0	9118 30.28 43.24 80.49	11971 39.75 56.76 63.73	21089 70.04
1	2210 7.34 24.49 19.51	6813 22.63 75.51 36.27	9023 29.96
Total	11328 37.62	18784 62.38	30112 100.00

APPENDIX A4: Frequency procedure by parity (2001 - 2010)

Table of HIVclass by Parity 2001			
HIVclass(HIVclass)	**Parity(Parity)**		
	-1	**1**	**Total**
0	4246 34.66 44.66 77.92	5262 42.96 55.34 77.37	9508 77.62
1	1203 9.82 43.87 22.08	1539 12.56 56.13 22.63	2742 22.38
Total	5449 44.48	6801 55.52	12250 100.00

Table of HIVclass by Parity 2002			
HIVclass(HIVclass)	**Parity(Parity)**		
	-1	**1**	**Total**
0	4775 31.14 42.42 77.84	6481 42.27 57.58 70.45	11256 73.41
1	1359 8.86 33.33 22.16	2718 17.73 66.67 29.55	4077 26.59
Total	6134 40.01	9199 59.99	15333 100.00

Table of HIVclass by Parity 2003			
HIVclass(HIVclass)	Parity(Parity)		
	-1	1	Total
0	607 32.46 44.44 72.35	759 40.59 55.56 73.62	1366 73.05
1	232 12.41 46.03 27.65	272 14.55 53.97 26.38	504 26.95
Total	839 44.87	1031 55.13	1870 100.00

Table of HIVclass by Parity 2004			
HIVclass(HIVclass)	Parity(Parity)		
	-1	1	Total
0	4581 30.92 43.82 76.01	5873 39.64 56.18 66.82	10454 70.56
1	1446 9.76 33.15 23.99	2916 19.68 66.85 33.18	4362 29.44
Total	6027 40.68	8789 59.32	14816 100.00

Table of HIVclass by Parity 2005			
HIVclass(HIVclass)	Parity(Parity)		
	-1	1	Total
0	4198 32.82 46.45 76.82	4839 37.83 53.55 66.06	9037 70.66
1	1267 9.91 33.76 23.18	2486 19.44 66.24 33.94	3753 29.34
Total	5465 42.73	7325 57.27	12790 100.00

Table of HIVclass by Parity 2006			
HIVclass(HIVclass)	Parity(Parity)		
	-1	1	Total
0	10301 32.95 46.26 78.05	11968 38.29 53.74 66.27	22269 71.24
1	2897 9.27 32.23 21.95	6092 19.49 67.77 33.73	8989 28.76
Total	13198 42.22	18060 57.78	31258 100.00

Table of HIVclass by Parity 2007			
HIVclass(HIVclass)	**Parity(Parity)**		
	-1	**1**	**Total**
0	10566 32.53 45.82 78.74	12493 38.46 54.18 65.54	23059 70.99
1	2853 8.78 30.28 21.26	6569 20.22 69.72 34.46	9422 29.01
Total	13419 41.31	19062 58.69	32481 100.00

Table of HIVclass by Parity 2008			
HIVclass(HIVclass)	**Parity(Parity)**		
	-1	**1**	**Total**
0	10876 33.39 47.08 78.91	12225 37.53 52.92 65.06	23101 70.93
1	2906 8.92 30.69 21.09	6564 20.15 69.31 34.94	9470 29.07
Total	13782 42.31	18789 57.69	32571 100.00

Table of HIVclass by Parity 2009			
HIVclass(HIVclass)	**Parity(Parity)**		
	-1	**1**	**Total**
0	10060 32.30 45.71 79.11	11949 38.37 54.29 64.85	22009 70.67
1	2656 8.53 29.08 20.89	6478 20.80 70.92 35.15	9134 29.33
Total	12716 40.83	18427 59.17	31143 100.00

Table of HIVclass by Parity 2010			
HIVclass(HIVclass)	**Parity(Parity)**		
	-1	**1**	**Total**
0	9669 32.11 45.85 79.18	11420 37.93 54.15 63.80	21089 70.04
1	2542 8.44 28.17 20.82	6481 21.52 71.83 36.20	9023 29.96
Total	12211 40.55	17901 59.45	30112 100.00

APPENDIX A5: Frequency procedure by education

Table of HIVclass by Education 2001			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	246	9262	9508
	2.01	75.61	77.62
	2.59	97.41	
	72.14	77.77	
1	95	2647	2742
	0.78	21.61	22.38
	3.46	96.54	
	27.86	22.23	
Total	341	11909	12250
	2.78	97.22	100.00

Table of HIVclass by Education 2002			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	4981	6275	11256
	32.49	40.92	73.41
	44.25	55.75	
	76.84	70.90	
1	1501	2576	4077
	9.79	16.80	26.59
	36.82	63.18	
	23.16	29.10	
Total	6482	8851	15333
	42.27	57.73	100.00

Table of HIVclass by Education 2003			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	17	1349	1366
	0.91	72.14	73.05
	1.24	98.76	
	85.00	72.92	
1	3	501	504
	0.16	26.79	26.95
	0.60	99.40	
	15.00	27.08	
Total	20	1850	1870
	1.07	98.93	100.00

Table of HIVclass by Education 2004			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	215	10239	10454
	1.45	69.11	70.56
	2.06	97.94	
	70.72	70.56	
1	89	4273	4362
	0.60	28.84	29.44
	2.04	97.96	
	29.28	29.44	
Total	304	14512	14816
	2.05	97.95	100.00

Table of HIVclass by Education 2005			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	163	8874	9037
	1.27	69.38	70.66
	1.80	98.20	
	71.49	70.64	
1	65	3688	3753
	0.51	28.84	29.34
	1.73	98.27	
	28.51	29.36	
Total	228	12562	12790
	1.78	98.22	100.00

Table of HIVclass by Education 2006			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	358	21911	22269
	1.15	70.10	71.24
	1.61	98.39	
	68.45	71.29	
1	165	8824	8989
	0.53	28.23	28.76
	1.84	98.16	
	31.55	28.71	
Total	523	30735	31258
	1.67	98.33	100.00

Table of HIVclass by Education 2007			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	725	22334	23059
	2.23	68.76	70.99
	3.14	96.86	
	68.79	71.07	
1	329	9093	9422
	1.01	27.99	29.01
	3.49	96.51	
	31.21	28.93	
Total	1054	31427	32481
	3.24	96.76	100.00

Table of HIVclass by Education 2008			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	226	22875	23101
	0.69	70.23	70.93
	0.98	99.02	
	62.26	71.02	
1	137	9333	9470
	0.42	28.65	29.07
	1.45	98.55	
	37.74	28.98	
Total	363	32208	32571
	1.11	98.89	100.00

Table of HIVclass by Education 2009			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	400	21609	22009
	1.28	69.39	70.67
	1.82	98.18	
	70.55	70.67	
1	167	8967	9134
	0.54	28.79	29.33
	1.83	98.17	
	29.45	29.33	
Total	567	30576	31143
	1.82	98.18	100.00

Table of HIVclass by Education 2010			
HIVclass(HIVclass)	Education(Education)		
	-1	1	Total
0	311	20778	21089
	1.03	69.00	70.04
	1.47	98.53	
	65.47	70.11	
1	164	8859	9023
	0.54	29.42	29.96
	1.82	98.18	
	34.53	29.89	
Total	475	29637	30112
	1.58	98.42	100.00

APPENDIX A6: Frequency procedure by syphilis (2001 - 2010)

Table of HIVclass by Syphilis 2001			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	9217	291	9508
	75.24	2.38	77.62
	96.94	3.06	
	77.78	72.75	
1	2633	109	2742
	21.49	0.89	22.38
	96.02	3.98	
	22.22	27.25	
Total	11850	400	12250
	96.73	3.27	100.00

Table of HIVclass by Syphilis 2002			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	10962 71.49 97.39 73.97	294 1.92 2.61 57.20	11256 73.41
1	3857 25.15 94.60 26.03	220 1.43 5.40 42.80	4077 26.59
Total	14819 96.65	514 3.35	15333 100.00

Table of HIVclass by Syphilis 2003			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	1321 70.64 96.71 73.39	45 2.41 3.29 64.29	1366 73.05
1	479 25.61 95.04 26.61	25 1.34 4.96 35.71	504 26.95
Total	1800 96.26	70 3.74	1870 100.00

Table of HIVclass by Syphilis 2004			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	10252	202	10454
	69.20	1.36	70.56
	98.07	1.93	
	70.82	59.41	
1	4224	138	4362
	28.51	0.93	29.44
	96.84	3.16	
	29.18	40.59	
Total	14476	340	14816
	97.71	2.29	100.00

Table of HIVclass by Syphilis 2005			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	8840	197	9037
	69.12	1.54	70.66
	97.82	2.18	
	70.85	62.94	
1	3637	116	3753
	28.44	0.91	29.34
	96.91	3.09	
	29.15	37.06	
Total	12477	313	12790
	97.55	2.45	100.00

Table of HIVclass by Syphilis 2006			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	21897	372	22269
	70.05	1.19	71.24
	98.33	1.67	
	71.40	63.27	
1	8773	216	8989
	28.07	0.69	28.76
	97.60	2.40	
	28.60	36.73	
Total	30670	588	31258
	98.12	1.88	100.00

Table of HIVclass by Syphilis 2007			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	22502	557	23059
	69.28	1.71	70.99
	97.58	2.42	
	71.23	62.44	
1	9087	335	9422
	27.98	1.03	29.01
	96.44	3.56	
	28.77	37.56	
Total	31589	892	32481
	97.25	2.75	100.00

Table of HIVclass by Syphilis 2008			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	22710	391	23101
	69.72	1.20	70.93
	98.31	1.69	
	71.08	62.86	
1	9239	231	9470
	28.37	0.71	29.07
	97.56	2.44	
	28.92	37.14	
Total	31949	622	32571
	98.09	1.91	100.00

Table of HIVclass by Syphilis 2009			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	21654	355	22009
	69.53	1.14	70.67
	98.39	1.61	
	70.90	59.17	
1	8889	245	9134
	28.54	0.79	29.33
	97.32	2.68	
	29.10	40.83	
Total	30543	600	31143
	98.07	1.93	100.00

Table of HIVclass by Syphilis 2010			
HIVclass(HIVclass)	Syphilis(Syphilis)		
	-1	1	Total
0	20838 69.20 98.81 70.24	251 0.83 1.19 56.15	21089 70.04
1	8827 29.31 97.83 29.76	196 0.65 2.17 43.85	9023 29.96
Total	29665 98.52	447 1.48	30112 100.00

APPENDIX B: TWO-LEVEL FULL FACTORIAL MODEL INPUT 2001 TO 2010

The two-level full factorial models have missing input values.

Appendix B.1: Two-level full factorial model input, 2001

RUN	MOTHAGE	FATHAGE	GRAV	PARITY	EDU	SYPH	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.26087
2	1	-1	-1	-1	-1	-1	.
3	-1	1	-1	-1	-1	-1	0.33333
4	1	1	-1	-1	-1	-1	0.27273
5	-1	-1	1	-1	-1	-1	0.75
6	1	-1	1	-1	-1	-1	1
7	-1	1	1	-1	-1	-1	1
8	1	1	1	-1	-1	-1	0.5
9	-1	-1	-1	1	-1	-1	.
10	1	-1	-1	1	-1	-1	.
11	-1	1	-1	1	-1	-1	1
12	1	1	-1	1	-1	-1	.
13	-1	-1	1	1	-1	-1	1
14	1	-1	1	1	-1	-1	0.33333
15	-1	1	1	1	-1	-1	0.22581
16	1	1	1	1	-1	-1	0.287
17	-1	-1	-1	-1	1	-1	0.14913
18	1	-1	-1	-1	1	-1	0.35714
19	-1	1	-1	-1	1	-1	0.28693
20	1	1	-1	-1	1	-1	0.26607
21	-1	-1	1	-1	1	-1	0.25472
22	1	-1	1	-1	1	-1	0.6
23	-1	1	1	-1	1	-1	0.24444
24	1	1	1	-1	1	-1	.
25	-1	-1	-1	1	1	-1	0.375
26	1	-1	-1	1	1	-1	0.5
27	-1	1	-1	1	1	-1	0.21053
28	1	1	-1	1	1	-1	0.23256
29	-1	-1	1	1	1	-1	0.21448
30	1	-1	1	1	1	-1	0.34615
31	-1	1	1	1	1	-1	0.30957
32	1	1	1	1	1	-1	0.19771
33	-1	-1	-1	-1	-1	1	.
34	1	-1	-1	-1	-1	1	.
35	-1	1	-1	-1	-1	1	.
36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	.
38	1	-1	1	-1	-1	1	.
39	-1	1	1	-1	-1	1	.

40	1	1	1	-1	-1	1	1
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.
43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	.
45	-1	-1	1	1	-1	1	1
46	1	-1	1	1	-1	1	.
47	-1	1	1	1	-1	1	0.5
48	1	1	1	1	-1	1	0.23362
49	-1	-1	-1	-1	1	1	0.16176
50	1	-1	-1	-1	1	1	0.5
51	-1	1	-1	-1	1	1	0.18
52	1	1	-1	-1	1	1	0.16667
53	-1	-1	1	-1	1	1	0.75
54	1	-1	1	-1	1	1	.
55	-1	1	1	-1	1	1	0.23077
56	1	1	1	-1	1	1	0.13333
57	-1	-1	-1	1	1	1	.
58	1	-1	-1	1	1	1	.
59	-1	1	-1	1	1	1	.
60	1	1	-1	1	1	1	0.33333
61	-1	-1	1	1	1	1	0.44444
62	1	-1	1	1	1	1	0.28571
63	-1	1	1	1	1	1	0.2973
64	1	1	1	1	1	1	0.32298

Appendix B.2: Two-level full factorial model input, 2002

RUN	MOTHAGE	FATHAGE	GRAVIDIT	PARITY	EDU	SYPHILIS	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.17241
2	1	-1	-1	-1	-1	-1	1
3	-1	1	-1	-1	-1	-1	0.25
4	1	1	-1	-1	-1	-1	0.26667
5	-1	-1	1	-1	-1	-1	0.1771
6	1	-1	1	-1	-1	-1	0.31154
7	-1	1	1	-1	-1	-1	0.30853
8	1	1	1	-1	-1	-1	0.34064
9	-1	-1	-1	1	-1	-1	.
10	1	-1	-1	1	-1	-1	1
11	-1	1	-1	1	-1	-1	.
12	1	1	-1	1	-1	-1	0.33333
13	-1	-1	1	1	-1	-1	0.32222
14	1	-1	1	1	-1	-1	0.36364
15	-1	1	1	1	-1	-1	0.3913

16	1	1	1	1	-1	-1	0.43846
17	-1	-1	-1	-1	1	-1	.
18	1	-1	-1	-1	1	-1	.
19	-1	1	-1	-1	1	-1	.
20	1	1	-1	-1	1	-1	1
21	-1	-1	1	-1	1	-1	0.12195
22	1	-1	1	-1	1	-1	0.42857
23	-1	1	1	-1	1	-1	0.33333
24	1	1	1	-1	1	-1	0.34483
25	-1	-1	-1	1	1	-1	0.25
26	1	-1	-1	1	1	-1	0.5
27	-1	1	-1	1	1	-1	0.46154
28	1	1	-1	1	1	-1	0.24828
29	-1	-1	1	1	1	-1	0.26342
30	1	-1	1	1	1	-1	0.33549
31	-1	1	1	1	1	-1	0.34712
32	1	1	1	1	1	-1	0.27994
33	-1	-1	-1	-1	-1	1	.
34	1	-1	-1	-1	-1	1	.
35	-1	1	-1	-1	-1	1	.
36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	0.36752
38	1	-1	1	-1	-1	1	0.5
39	-1	1	1	-1	-1	1	0.625
40	1	1	1	-1	-1	1	0.45455
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.
43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	.
45	-1	-1	1	1	-1	1	0.58333
46	1	-1	1	1	-1	1	1
47	-1	1	1	1	-1	1	0.75
48	1	1	1	1	-1	1	0.4
49	-1	-1	-1	-1	1	1	.
50	1	-1	-1	-1	1	1	.
51	-1	1	-1	-1	1	1	1
52	1	1	-1	-1	1	1	.
53	-1	-1	1	-1	1	1	1
54	1	-1	1	-1	1	1	.
55	-1	1	1	-1	1	1	.
56	1	1	1	-1	1	1	0.25
57	-1	-1	-1	1	1	1	1
58	1	-1	-1	1	1	1	.
59	-1	1	-1	1	1	1	1

60	1	1	-1	1	1	1	0.18182
61	-1	-1	1	1	1	1	0.48148
62	1	-1	1	1	1	1	0.73077
63	-1	1	1	1	1	1	0.48
64	1	1	1	1	1	1	0.36979

Appendix B.3: Two-level full factorial model input, 2003

RUN	MOTHAGE	FATHAGE	GRAV	PARITY	EDU	SYPH	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.26087
2	1	-1	-1	-1	-1	-1	.
3	-1	1	-1	-1	-1	-1	0.33333
4	1	1	-1	-1	-1	-1	0.27273
5	-1	-1	1	-1	-1	-1	0.75
6	1	-1	1	-1	-1	-1	1
7	-1	1	1	-1	-1	-1	1
8	1	1	1	-1	-1	-1	0.5
9	-1	-1	-1	1	-1	-1	.
10	1	-1	-1	1	-1	-1	.
11	-1	1	-1	1	-1	-1	1
12	1	1	-1	1	-1	-1	.
13	-1	-1	1	1	-1	-1	1
14	1	-1	1	1	-1	-1	0.33333
15	-1	1	1	1	-1	-1	0.22581
16	1	1	1	1	-1	-1	0.287
17	-1	-1	-1	-1	1	-1	0.14913
18	1	-1	-1	-1	1	-1	0.35714
19	-1	1	-1	-1	1	-1	0.28693
20	1	1	-1	-1	1	-1	0.26607
21	-1	-1	1	-1	1	-1	0.25472
22	1	-1	1	-1	1	-1	0.6
23	-1	1	1	-1	1	-1	0.24444
24	1	1	1	-1	1	-1	.
25	-1	-1	-1	1	1	-1	0.375
26	1	-1	-1	1	1	-1	0.5
27	-1	1	-1	1	1	-1	0.21053
28	1	1	-1	1	1	-1	0.23256
29	-1	-1	1	1	1	-1	0.21448
30	1	-1	1	1	1	-1	0.34615
31	-1	1	1	1	1	-1	0.30957
32	1	1	1	1	1	-1	0.19771
33	-1	-1	-1	-1	-1	1	.
34	1	-1	-1	-1	-1	1	.
35	-1	1	-1	-1	-1	1	.

36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	.
38	1	-1	1	-1	-1	1	.
39	-1	1	1	-1	-1	1	.
40	1	1	1	-1	-1	1	1
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.
43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	.
45	-1	-1	1	1	-1	1	1
46	1	-1	1	1	-1	1	.
47	-1	1	1	1	-1	1	0.5
48	1	1	1	1	-1	1	0.23362
49	-1	-1	-1	-1	1	1	0.16176
50	1	-1	-1	-1	1	1	0.5
51	-1	1	-1	-1	1	1	0.18
52	1	1	-1	-1	1	1	0.16667
53	-1	-1	1	-1	1	1	0.75
54	1	-1	1	-1	1	1	.
55	-1	1	1	-1	1	1	0.23077
56	1	1	1	-1	1	1	0.13333
57	-1	-1	-1	1	1	1	.
58	1	-1	-1	1	1	1	.
59	-1	1	-1	1	1	1	.
60	1	1	-1	1	1	1	0.33333
61	-1	-1	1	1	1	1	0.44444
62	1	-1	1	1	1	1	0.28571
63	-1	1	1	1	1	1	0.2973
64	1	1	1	1	1	1	0.32298

Appendix B.4: Two level full factorial model input, 2004

RUN	MOTHAGE	FATHAGE	GRAV	PARITY	EDU	SYPH	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.26087
2	1	-1	-1	-1	-1	-1	.
3	-1	1	-1	-1	-1	-1	0.33333
4	1	1	-1	-1	-1	-1	0.27273
5	-1	-1	1	-1	-1	-1	0.75
6	1	-1	1	-1	-1	-1	1
7	-1	1	1	-1	-1	-1	1
8	1	1	1	-1	-1	-1	0.5
9	-1	-1	-1	1	-1	-1	.
10	1	-1	-1	1	-1	-1	.
11	-1	1	-1	1	-1	-1	1
12	1	1	-1	1	-1	-1	.
13	-1	-1	1	1	-1	-1	1
14	1	-1	1	1	-1	-1	0.33333
15	-1	1	1	1	-1	-1	0.22581
16	1	1	1	1	-1	-1	0.287
17	-1	-1	-1	-1	1	-1	0.14913
18	1	-1	-1	-1	1	-1	0.35714
19	-1	1	-1	-1	1	-1	0.28693
20	1	1	-1	-1	1	-1	0.26607
21	-1	-1	1	-1	1	-1	0.25472
22	1	-1	1	-1	1	-1	0.6
23	-1	1	1	-1	1	-1	0.24444
24	1	1	1	-1	1	-1	.
25	-1	-1	-1	1	1	-1	0.375
26	1	-1	-1	1	1	-1	0.5
27	-1	1	-1	1	1	-1	0.21053
28	1	1	-1	1	1	-1	0.23256
29	-1	-1	1	1	1	-1	0.21448
30	1	-1	1	1	1	-1	0.34615
31	-1	1	1	1	1	-1	0.30957
32	1	1	1	1	1	-1	0.19771
33	-1	-1	-1	-1	-1	1	.
34	1	-1	-1	-1	-1	1	.
35	-1	1	-1	-1	-1	1	.
36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	.
38	1	-1	1	-1	-1	1	.
39	-1	1	1	-1	-1	1	.
40	1	1	1	-1	-1	1	1
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.

43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	.
45	-1	-1	1	1	-1	1	1
46	1	-1	1	1	-1	1	.
47	-1	1	1	1	-1	1	0.5
48	1	1	1	1	-1	1	0.23362
49	-1	-1	-1	-1	1	1	0.16176
50	1	-1	-1	-1	1	1	0.5
51	-1	1	-1	-1	1	1	0.18
52	1	1	-1	-1	1	1	0.16667
53	-1	-1	1	-1	1	1	0.75
54	1	-1	1	-1	1	1	.
55	-1	1	1	-1	1	1	0.23077
56	1	1	1	-1	1	1	0.13333
57	-1	-1	-1	1	1	1	.
58	1	-1	-1	1	1	1	.
59	-1	1	-1	1	1	1	.
60	1	1	-1	1	1	1	0.33333
61	-1	-1	1	1	1	1	0.44444
62	1	-1	1	1	1	1	0.28571
63	-1	1	1	1	1	1	0.2973
64	1	1	1	1	1	1	0.32298

Appendix B.5: Two-level full factorial model input, 2005

RUN	MOTHAGE	FATHAGE	GRAV	PARITY	EDU	SYPH	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.26087
2	1	-1	-1	-1	-1	-1	.
3	-1	1	-1	-1	-1	-1	0.33333
4	1	1	-1	-1	-1	-1	0.27273
5	-1	-1	1	-1	-1	-1	0.75
6	1	-1	1	-1	-1	-1	1
7	-1	1	1	-1	-1	-1	1
8	1	1	1	-1	-1	-1	0.5
9	-1	-1	-1	1	-1	-1	.
10	1	-1	-1	1	-1	-1	.
11	-1	1	-1	1	-1	-1	1
12	1	1	-1	1	-1	-1	.
13	-1	-1	1	1	-1	-1	1
14	1	-1	1	1	-1	-1	0.33333
15	-1	1	1	1	-1	-1	0.22581
16	1	1	1	1	-1	-1	0.287
17	-1	-1	-1	-1	1	-1	0.14913
18	1	-1	-1	-1	1	-1	0.35714

19	-1	1	-1	-1	1	-1	0.28693
20	1	1	-1	-1	1	-1	0.26607
21	-1	-1	1	-1	1	-1	0.25472
22	1	-1	1	-1	1	-1	0.6
23	-1	1	1	-1	1	-1	0.24444
24	1	1	1	-1	1	-1	.
25	-1	-1	-1	1	1	-1	0.375
26	1	-1	-1	1	1	-1	0.5
27	-1	1	-1	1	1	-1	0.21053
28	1	1	-1	1	1	-1	0.23256
29	-1	-1	1	1	1	-1	0.21448
30	1	-1	1	1	1	-1	0.34615
31	-1	1	1	1	1	-1	0.30957
32	1	1	1	1	1	-1	0.19771
33	-1	-1	-1	-1	-1	1	.
34	1	-1	-1	-1	-1	1	.
35	-1	1	-1	-1	-1	1	.
36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	.
38	1	-1	1	-1	-1	1	.
39	-1	1	1	-1	-1	1	.
40	1	1	1	-1	-1	1	1
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.
43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	.
45	-1	-1	1	1	-1	1	1
46	1	-1	1	1	-1	1	.
47	-1	1	1	1	-1	1	0.5
48	1	1	1	1	-1	1	0.23362
49	-1	-1	-1	-1	1	1	0.16176
50	1	-1	-1	-1	1	1	0.5
51	-1	1	-1	-1	1	1	0.18
52	1	1	-1	-1	1	1	0.16667
53	-1	-1	1	-1	1	1	0.75
54	1	-1	1	-1	1	1	.
55	-1	1	1	-1	1	1	0.23077
56	1	1	1	-1	1	1	0.13333
57	-1	-1	-1	1	1	1	.
58	1	-1	-1	1	1	1	.
59	-1	1	-1	1	1	1	.
60	1	1	-1	1	1	1	0.33333
61	-1	-1	1	1	1	1	0.44444
62	1	-1	1	1	1	1	0.28571

63	-1	1	1	1	1	1	0.2973
64	1	1	1	1	1	1	0.32298

Appendix B.6: Two-level full factorial model input, 2006

RUN	MOTHAGE	FATHAGE	GRAV	PARITY	EDU	SYPH	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.26087
2	1	-1	-1	-1	-1	-1	.
3	-1	1	-1	-1	-1	-1	0.33333
4	1	1	-1	-1	-1	-1	0.27273
5	-1	-1	1	-1	-1	-1	0.75
6	1	-1	1	-1	-1	-1	1
7	-1	1	1	-1	-1	-1	1
8	1	1	1	-1	-1	-1	0.5
9	-1	-1	-1	1	-1	-1	.
10	1	-1	-1	1	-1	-1	.
11	-1	1	-1	1	-1	-1	1
12	1	1	-1	1	-1	-1	.
13	-1	-1	1	1	-1	-1	1
14	1	-1	1	1	-1	-1	0.33333
15	-1	1	1	1	-1	-1	0.22581
16	1	1	1	1	-1	-1	0.287
17	-1	-1	-1	-1	1	-1	0.14913
18	1	-1	-1	-1	1	-1	0.35714
19	-1	1	-1	-1	1	-1	0.28693
20	1	1	-1	-1	1	-1	0.26607
21	-1	-1	1	-1	1	-1	0.25472
22	1	-1	1	-1	1	-1	0.6
23	-1	1	1	-1	1	-1	0.24444
24	1	1	1	-1	1	-1	.
25	-1	-1	-1	1	1	-1	0.375
26	1	-1	-1	1	1	-1	0.5
27	-1	1	-1	1	1	-1	0.21053
28	1	1	-1	1	1	-1	0.23256
29	-1	-1	1	1	1	-1	0.21448
30	1	-1	1	1	1	-1	0.34615
31	-1	1	1	1	1	-1	0.30957
32	1	1	1	1	1	-1	0.19771
33	-1	-1	-1	-1	-1	1	.
34	1	-1	-1	-1	-1	1	.
35	-1	1	-1	-1	-1	1	.
36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	.
38	1	-1	1	-1	-1	1	.

39	-1	1	1	-1	-1	1	.
40	1	1	1	-1	-1	1	1
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.
43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	.
45	-1	-1	1	1	-1	1	1
46	1	-1	1	1	-1	1	.
47	-1	1	1	1	-1	1	0.5
48	1	1	1	1	-1	1	0.23362
49	-1	-1	-1	-1	1	1	0.16176
50	1	-1	-1	-1	1	1	0.5
51	-1	1	-1	-1	1	1	0.18
52	1	1	-1	-1	1	1	0.16667
53	-1	-1	1	-1	1	1	0.75
54	1	-1	1	-1	1	1	.
55	-1	1	1	-1	1	1	0.23077
56	1	1	1	-1	1	1	0.13333
57	-1	-1	-1	1	1	1	.
58	1	-1	-1	1	1	1	.
59	-1	1	-1	1	1	1	.
60	1	1	-1	1	1	1	0.33333
61	-1	-1	1	1	1	1	0.44444
62	1	-1	1	1	1	1	0.28571
63	-1	1	1	1	1	1	0.2973
64	1	1	1	1	1	1	0.32298

Appendix B.7: Two-level full factorial model input, 2007

RUN	MOTHAGE	FATHAGE	GRAV	PARITY	EDU	SYPH	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.26087
2	1	-1	-1	-1	-1	-1	.
3	-1	1	-1	-1	-1	-1	0.33333
4	1	1	-1	-1	-1	-1	0.27273
5	-1	-1	1	-1	-1	-1	0.75
6	1	-1	1	-1	-1	-1	1
7	-1	1	1	-1	-1	-1	1
8	1	1	1	-1	-1	-1	0.5
9	-1	-1	-1	1	-1	-1	.
10	1	-1	-1	1	-1	-1	.
11	-1	1	-1	1	-1	-1	1
12	1	1	-1	1	-1	-1	.
13	-1	-1	1	1	-1	-1	1
14	1	-1	1	1	-1	-1	0.33333

15	-1	1	1	1	-1	-1	0.22581
16	1	1	1	1	-1	-1	0.287
17	-1	-1	-1	-1	1	-1	0.14913
18	1	-1	-1	-1	1	-1	0.35714
19	-1	1	-1	-1	1	-1	0.28693
20	1	1	-1	-1	1	-1	0.26607
21	-1	-1	1	-1	1	-1	0.25472
22	1	-1	1	-1	1	-1	0.6
23	-1	1	1	-1	1	-1	0.24444
24	1	1	1	-1	1	-1	.
25	-1	-1	-1	1	1	-1	0.375
26	1	-1	-1	1	1	-1	0.5
27	-1	1	-1	1	1	-1	0.21053
28	1	1	-1	1	1	-1	0.23256
29	-1	-1	1	1	1	-1	0.21448
30	1	-1	1	1	1	-1	0.34615
31	-1	1	1	1	1	-1	0.30957
32	1	1	1	1	1	-1	0.19771
33	-1	-1	-1	-1	-1	1	.
34	1	-1	-1	-1	-1	1	.
35	-1	1	-1	-1	-1	1	.
36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	.
38	1	-1	1	-1	-1	1	.
39	-1	1	1	-1	-1	1	.
40	1	1	1	-1	-1	1	1
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.
43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	.
45	-1	-1	1	1	-1	1	1
46	1	-1	1	1	-1	1	.
47	-1	1	1	1	-1	1	0.5
48	1	1	1	1	-1	1	0.23362
49	-1	-1	-1	-1	1	1	0.16176
50	1	-1	-1	-1	1	1	0.5
51	-1	1	-1	-1	1	1	0.18
52	1	1	-1	-1	1	1	0.16667
53	-1	-1	1	-1	1	1	0.75
54	1	-1	1	-1	1	1	.
55	-1	1	1	-1	1	1	0.23077
56	1	1	1	-1	1	1	0.13333
57	-1	-1	-1	1	1	1	.
58	1	-1	-1	1	1	1	.

59	-1	1	-1	1	1	1	.
60	1	1	-1	1	1	1	0.33333
61	-1	-1	1	1	1	1	0.44444
62	1	-1	1	1	1	1	0.28571
63	-1	1	1	1	1	1	0.2973
64	1	1	1	1	1	1	0.32298

Appendix B.8: Two-level full factorial model input, 2008

RUN	MOTHAGE	FATHAGE	GRAV	PARITY	EDU	SYPH	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.26087
2	1	-1	-1	-1	-1	-1	.
3	-1	1	-1	-1	-1	-1	0.33333
4	1	1	-1	-1	-1	-1	0.27273
5	-1	-1	1	-1	-1	-1	0.75
6	1	-1	1	-1	-1	-1	1
7	-1	1	1	-1	-1	-1	1
8	1	1	1	-1	-1	-1	0.5
9	-1	-1	-1	1	-1	-1	.
10	1	-1	-1	1	-1	-1	.
11	-1	1	-1	1	-1	-1	1
12	1	1	-1	1	-1	-1	.
13	-1	-1	1	1	-1	-1	1
14	1	-1	1	1	-1	-1	0.33333
15	-1	1	1	1	-1	-1	0.22581
16	1	1	1	1	-1	-1	0.287
17	-1	-1	-1	-1	1	-1	0.14913
18	1	-1	-1	-1	1	-1	0.35714
19	-1	1	-1	-1	1	-1	0.28693
20	1	1	-1	-1	1	-1	0.26607
21	-1	-1	1	-1	1	-1	0.25472
22	1	-1	1	-1	1	-1	0.6
23	-1	1	1	-1	1	-1	0.24444
24	1	1	1	-1	1	-1	.
25	-1	-1	-1	1	1	-1	0.375
26	1	-1	-1	1	1	-1	0.5
27	-1	1	-1	1	1	-1	0.21053
28	1	1	-1	1	1	-1	0.23256
29	-1	-1	1	1	1	-1	0.21448
30	1	-1	1	1	1	-1	0.34615
31	-1	1	1	1	1	-1	0.30957
32	1	1	1	1	1	-1	0.19771
33	-1	-1	-1	-1	-1	1	.
34	1	-1	-1	-1	-1	1	.

35	-1	1	-1	-1	-1	1	.
36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	.
38	1	-1	1	-1	-1	1	.
39	-1	1	1	-1	-1	1	.
40	1	1	1	-1	-1	1	1
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.
43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	.
45	-1	-1	1	1	-1	1	1
46	1	-1	1	1	-1	1	.
47	-1	1	1	1	-1	1	0.5
48	1	1	1	1	-1	1	0.23362
49	-1	-1	-1	-1	1	1	0.16176
50	1	-1	-1	-1	1	1	0.5
51	-1	1	-1	-1	1	1	0.18
52	1	1	-1	-1	1	1	0.16667
53	-1	-1	1	-1	1	1	0.75
54	1	-1	1	-1	1	1	.
55	-1	1	1	-1	1	1	0.23077
56	1	1	1	-1	1	1	0.13333
57	-1	-1	-1	1	1	1	.
58	1	-1	-1	1	1	1	.
59	-1	1	-1	1	1	1	.
60	1	1	-1	1	1	1	0.33333
61	-1	-1	1	1	1	1	0.44444
62	1	-1	1	1	1	1	0.28571
63	-1	1	1	1	1	1	0.2973
64	1	1	1	1	1	1	0.32298

Appendix B.9: Two-level full factorial model input, 2009

RUN	MOTHAGE	FATHAGE	GRAVIDIT	PARITY	EDUCATIO	SYPHILIS	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.17778
2	1	-1	-1	-1	-1	-1	1
3	-1	1	-1	-1	-1	-1	0.22222
4	1	1	-1	-1	-1	-1	0.25
5	-1	-1	1	-1	-1	-1	0.25
6	1	-1	1	-1	-1	-1	.
7	-1	1	1	-1	-1	-1	1
8	1	1	1	-1	-1	-1	0.2
9	-1	-1	-1	1	-1	-1	1
10	1	-1	-1	1	-1	-1	1
11	-1	1	-1	1	-1	-1	1
12	1	1	-1	1	-1	-1	1
13	-1	-1	1	1	-1	-1	0.17949
14	1	-1	1	1	-1	-1	0.4
15	-1	1	1	1	-1	-1	0.42857
16	1	1	1	1	-1	-1	0.31579
17	-1	-1	-1	-1	1	-1	0.00385
18	1	-1	-1	-1	1	-1	0.26716
19	-1	1	-1	-1	1	-1	0.27982
20	1	1	-1	-1	1	-1	0.35029
21	-1	-1	1	-1	1	-1	0.26008
22	1	-1	1	-1	1	-1	0.43925
23	-1	1	1	-1	1	-1	0.34706
24	1	1	1	-1	1	-1	0.48903
25	-1	-1	-1	1	1	-1	0.15328
26	1	-1	-1	1	1	-1	0.36842
27	-1	1	-1	1	1	-1	0.30435
28	1	1	-1	1	1	-1	0.27451
29	-1	-1	1	1	1	-1	0.26657
30	1	-1	1	1	1	-1	0.34555
31	-1	1	1	1	1	-1	0.35724
32	1	1	1	1	1	-1	0.3798
33	-1	-1	-1	-1	-1	1	1
34	1	-1	-1	-1	-1	1	.
35	-1	1	-1	-1	-1	1	.
36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	1
38	1	-1	1	-1	-1	1	.
39	-1	1	1	-1	-1	1	.
40	1	1	1	-1	-1	1	.
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.

43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	1
45	-1	-1	1	1	-1	1	0.5
46	1	-1	1	1	-1	1	.
47	-1	1	1	1	-1	1	.
48	1	1	1	1	-1	1	0.30769
49	-1	-1	-1	-1	1	1	0.99568
50	1	-1	-1	-1	1	1	0.5
51	-1	1	-1	-1	1	1	0.31579
52	1	1	-1	-1	1	1	0.30769
53	-1	-1	1	-1	1	1	0.45455
54	1	-1	1	-1	1	1	1
55	-1	1	1	-1	1	1	0.71429
56	1	1	1	-1	1	1	0.41667
57	-1	-1	-1	1	1	1	1
58	1	-1	-1	1	1	1	1
59	-1	1	-1	1	1	1	.
60	1	1	-1	1	1	1	0.5
61	-1	-1	1	1	1	1	0.375
62	1	-1	1	1	1	1	0.60976
63	-1	1	1	1	1	1	0.34783
64	1	1	1	1	1	1	0.46983

Appendix B.10: Two-level full factorial model input, 2010

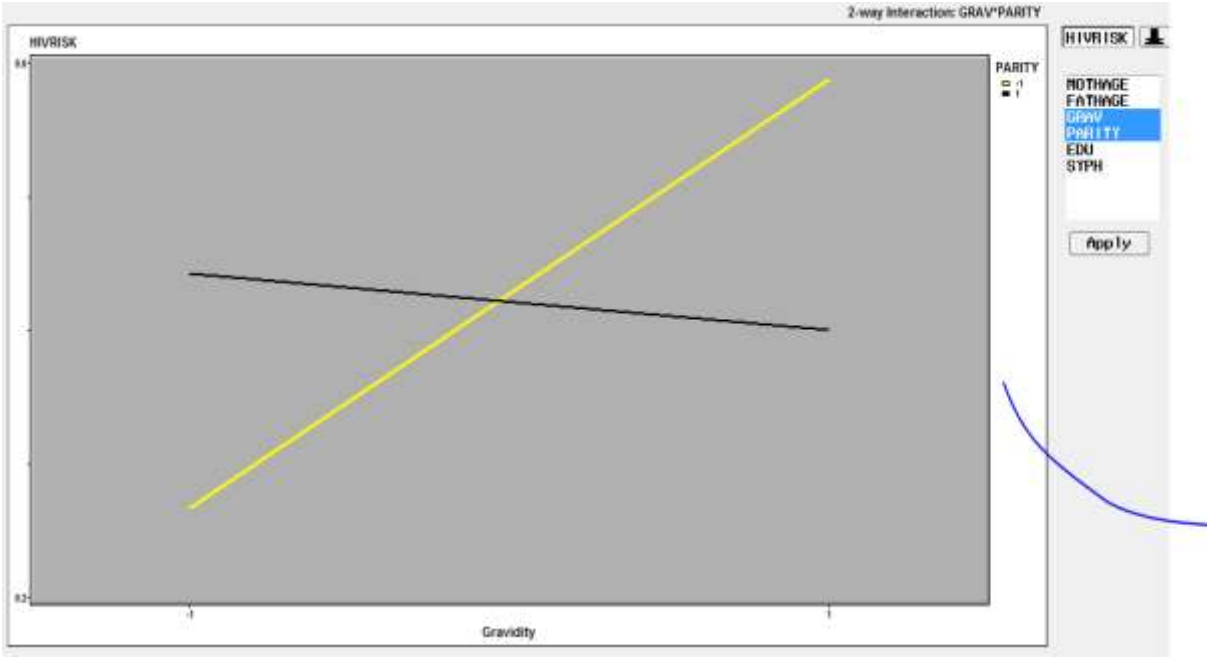
RUN	MOTHAGE	FATHAGE	GRAVIDIT	PARITY	EDUCATIO	SYPHILIS	HIVRISK
1	-1	-1	-1	-1	-1	-1	0.16327
2	1	-1	-1	-1	-1	-1	1
3	-1	1	-1	-1	-1	-1	0.2
4	1	1	-1	-1	-1	-1	0.25
5	-1	-1	1	-1	-1	-1	1
6	1	-1	1	-1	-1	-1	0.5
7	-1	1	1	-1	-1	-1	0.75
8	1	1	1	-1	-1	-1	0.25
9	-1	-1	-1	1	-1	-1	1
10	1	-1	-1	1	-1	-1	.
11	-1	1	-1	1	-1	-1	.
12	1	1	-1	1	-1	-1	1
13	-1	-1	1	1	-1	-1	0.33333
14	1	-1	1	1	-1	-1	0.30769
15	-1	1	1	1	-1	-1	0.26923
16	1	1	1	1	-1	-1	0.37895
17	-1	-1	-1	-1	1	-1	0.1525
18	1	-1	-1	-1	1	-1	0.26923

19	-1	1	-1	-1	1	-1	0.27719
20	1	1	-1	-1	1	-1	0.36229
21	-1	-1	1	-1	1	-1	0.2233
22	1	-1	1	-1	1	-1	0.36283
23	-1	1	1	-1	1	-1	0.33571
24	1	1	1	-1	1	-1	0.50262
25	-1	-1	-1	1	1	-1	0.12222
26	1	-1	-1	1	1	-1	0.45833
27	-1	1	-1	1	1	-1	0.26087
28	1	1	-1	1	1	-1	0.44118
29	-1	-1	1	1	1	-1	0.27135
30	1	-1	1	1	1	-1	0.35064
31	-1	1	1	1	1	-1	0.36442
32	1	1	1	1	1	-1	0.39361
33	-1	-1	-1	-1	-1	1	1
34	1	-1	-1	-1	-1	1	.
35	-1	1	-1	-1	-1	1	.
36	1	1	-1	-1	-1	1	.
37	-1	-1	1	-1	-1	1	.
38	1	-1	1	-1	-1	1	.
39	-1	1	1	-1	-1	1	.
40	1	1	1	-1	-1	1	.
41	-1	-1	-1	1	-1	1	.
42	1	-1	-1	1	-1	1	.
43	-1	1	-1	1	-1	1	.
44	1	1	-1	1	-1	1	.
45	-1	-1	1	1	-1	1	1
46	1	-1	1	1	-1	1	.
47	-1	1	1	1	-1	1	.
48	1	1	1	1	-1	1	0.33333
49	-1	-1	-1	-1	1	1	0.29091
50	1	-1	-1	-1	1	1	0.6
51	-1	1	-1	-1	1	1	0.2381
52	1	1	-1	-1	1	1	0.52941
53	-1	-1	1	-1	1	1	0.18182
54	1	-1	1	-1	1	1	1
55	-1	1	1	-1	1	1	1
56	1	1	1	-1	1	1	0.5
57	-1	-1	-1	1	1	1	1
58	1	-1	-1	1	1	1	.
59	-1	1	-1	1	1	1	.
60	1	1	-1	1	1	1	1
61	-1	-1	1	1	1	1	0.5
62	1	-1	1	1	1	1	0.42857

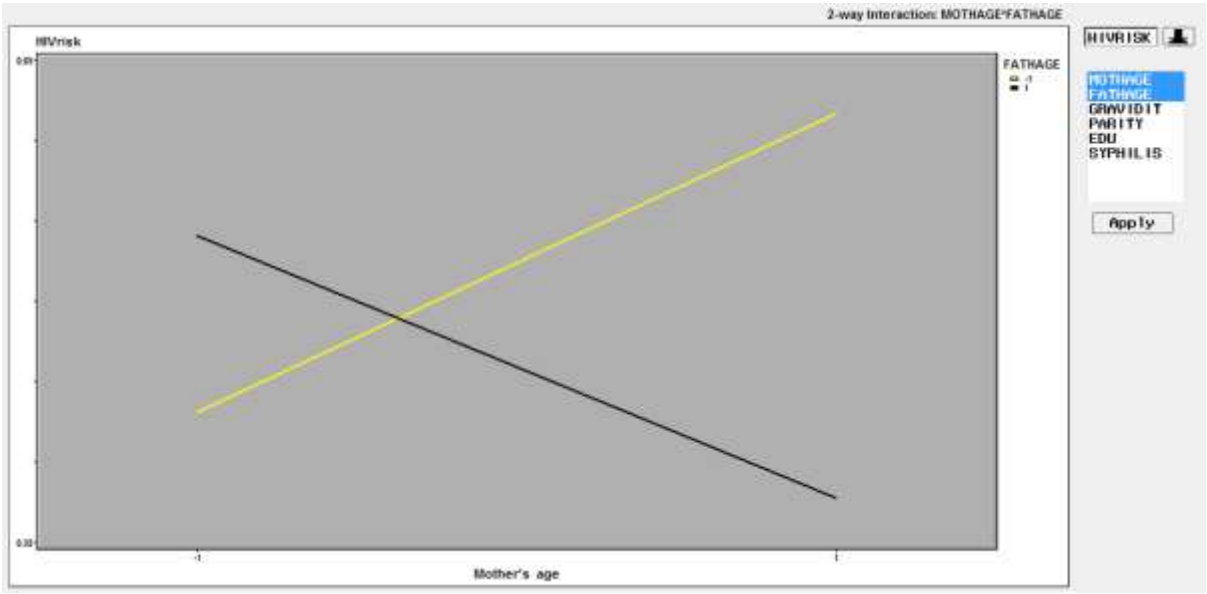
63	-1	1	1	1	1	1	0.63636
64	1	1	1	1	1	1	0.50667

APPENDIX C: INTERACTION PLOT

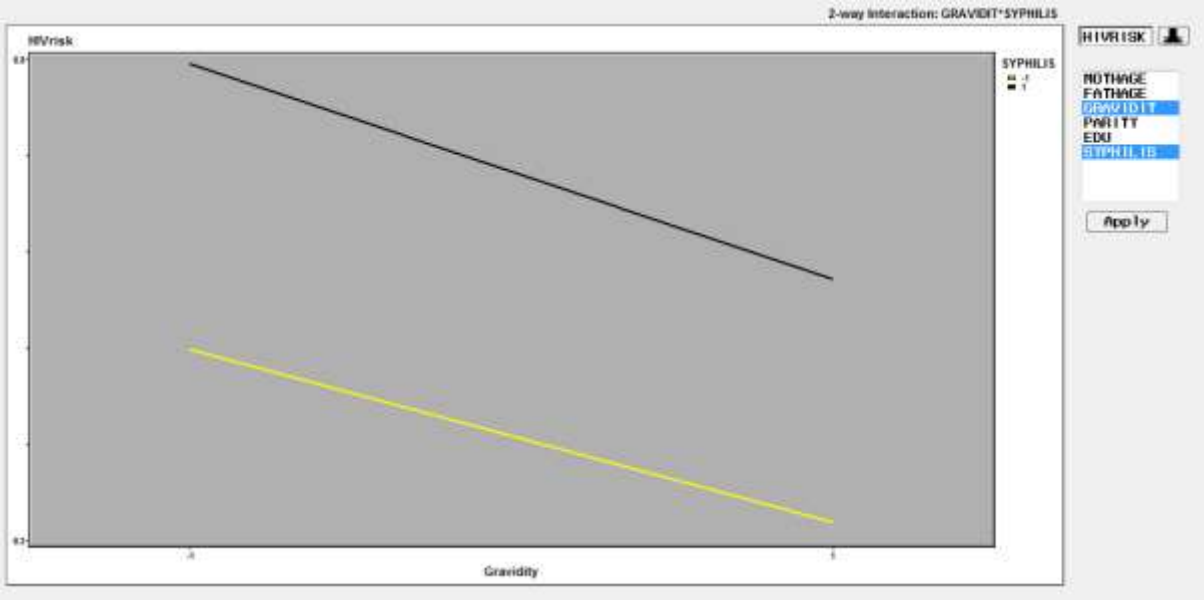
Interaction Plots 2001: Gravidity*Parity



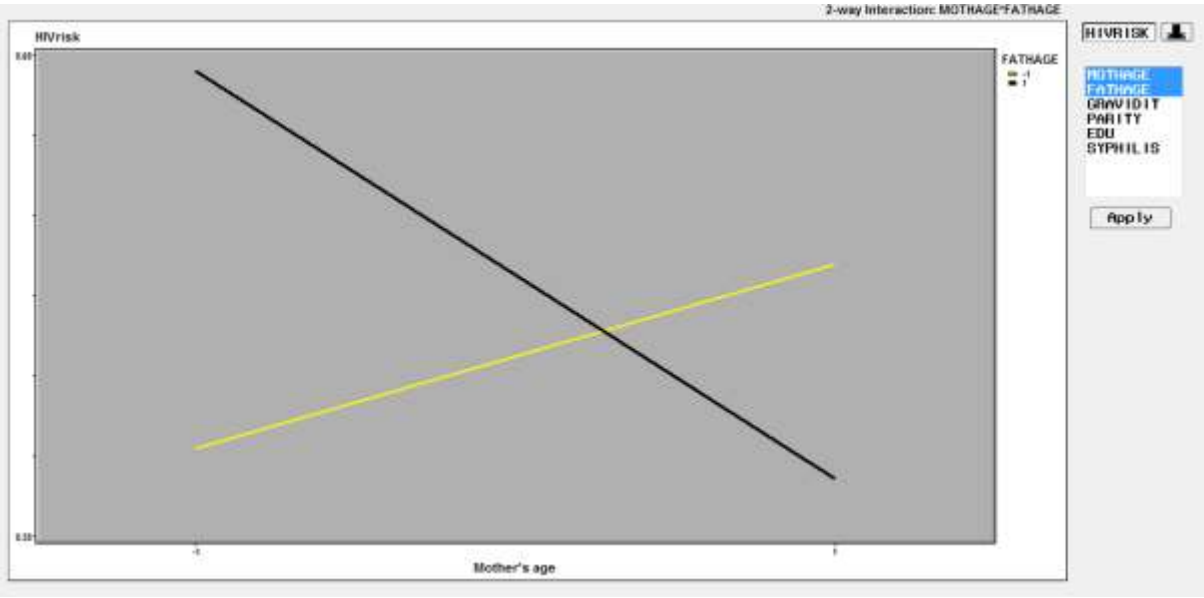
Interaction plot 2002: Mothers age* partners age



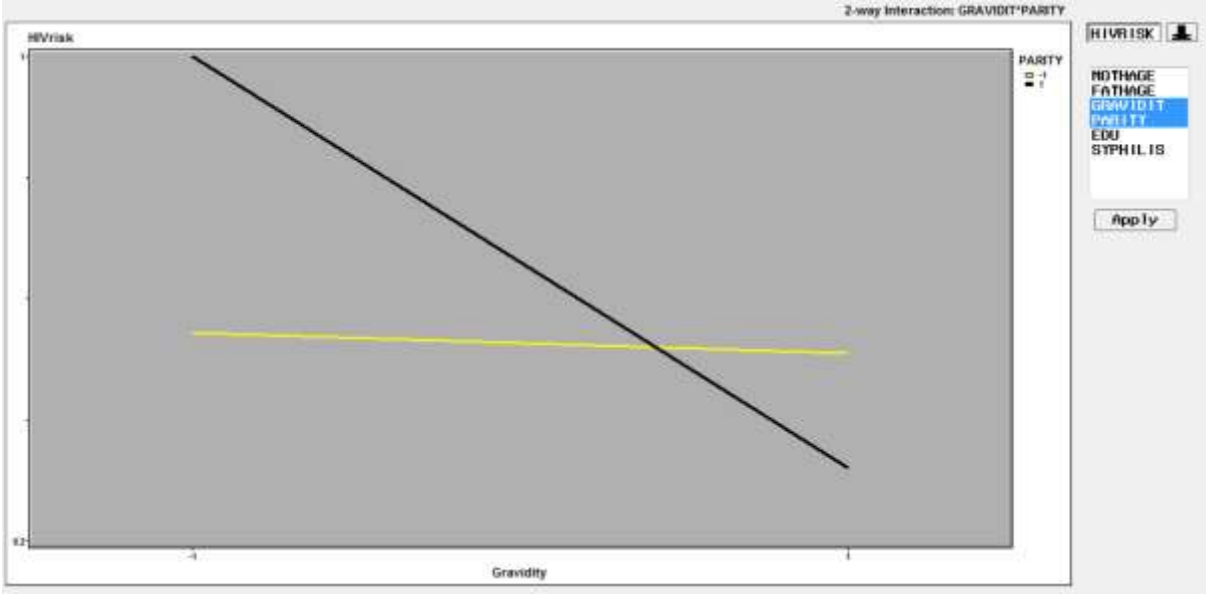
Interaction plot 2002: Gravidity * Syphilis



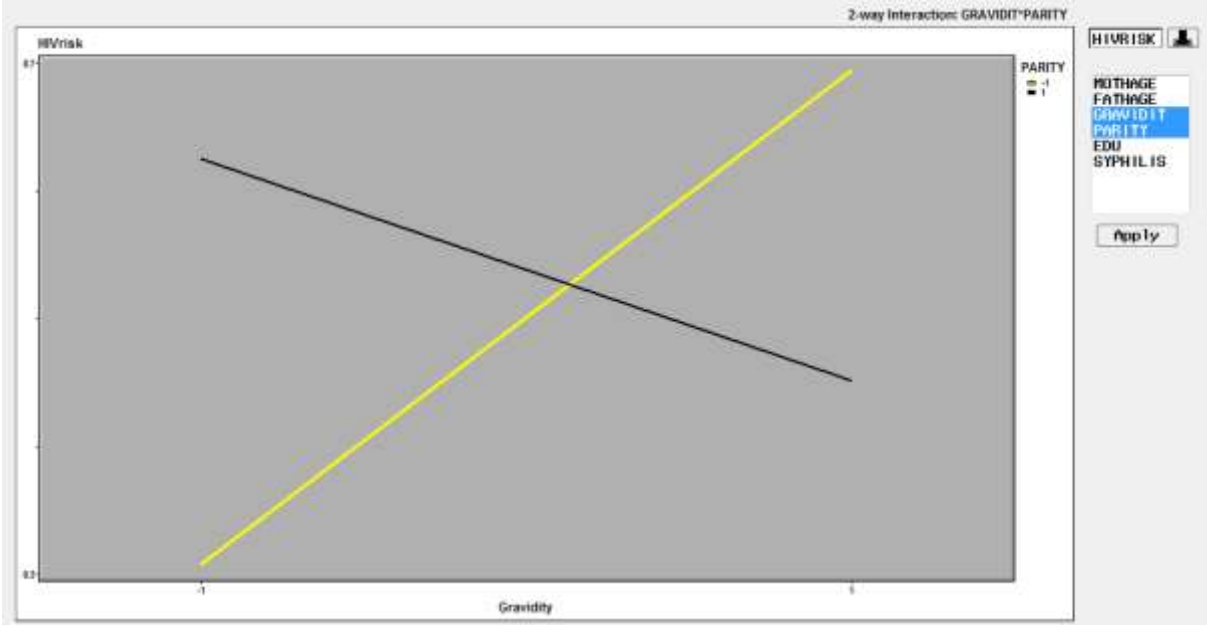
Interaction plot 2003: Mother's age * Partner's age



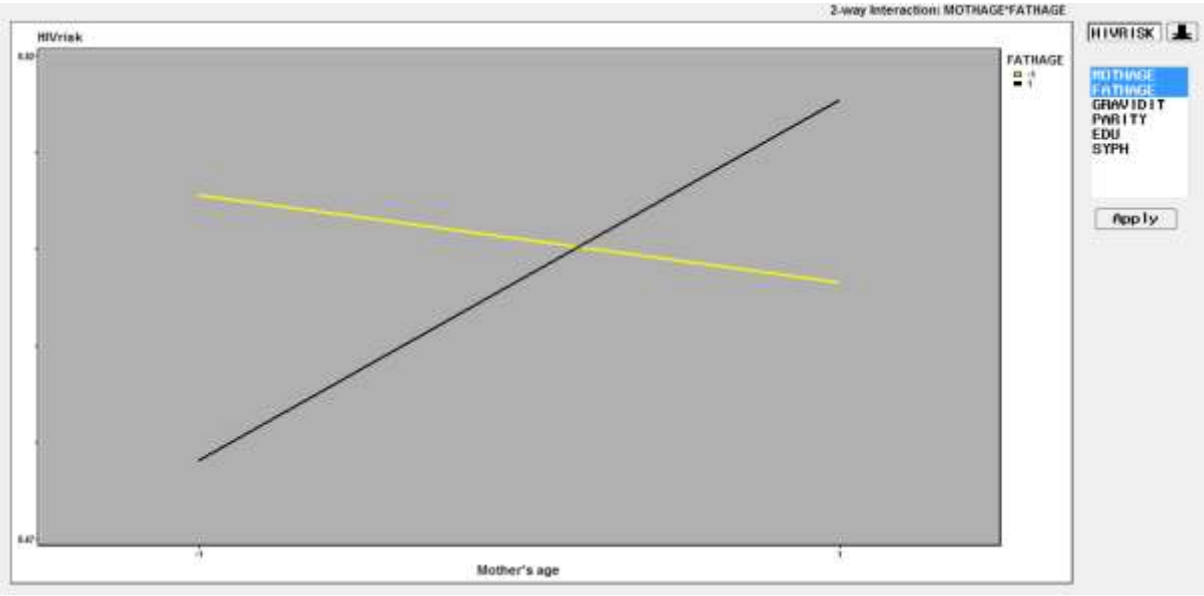
Interaction plot 2003: Gravidity * Parity



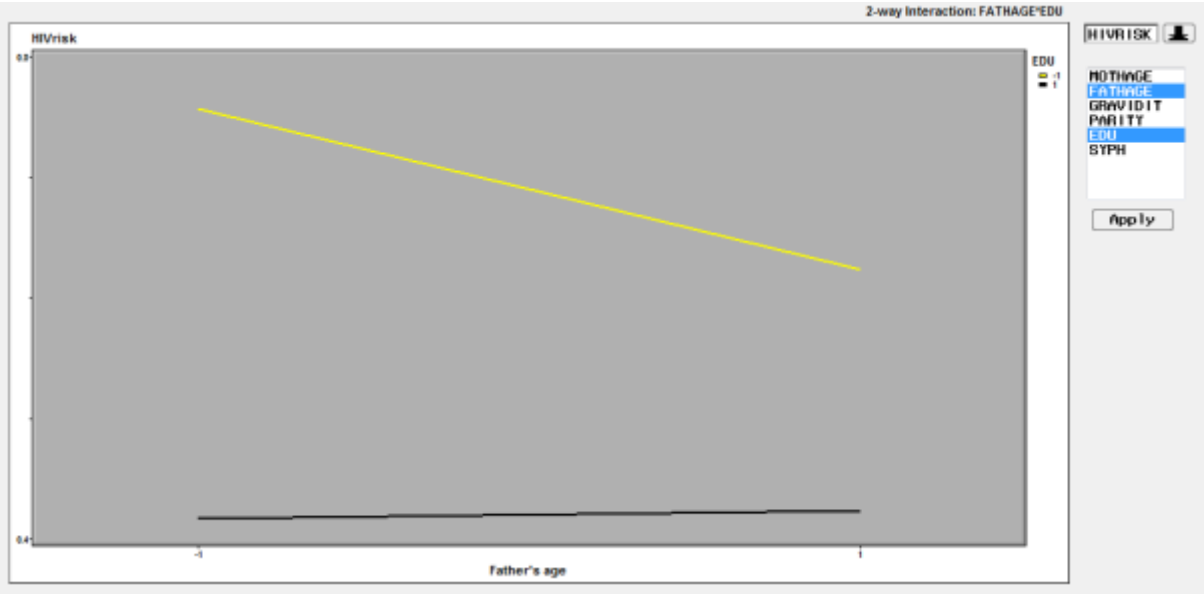
Interaction plot 2004: Gravidity * Parity



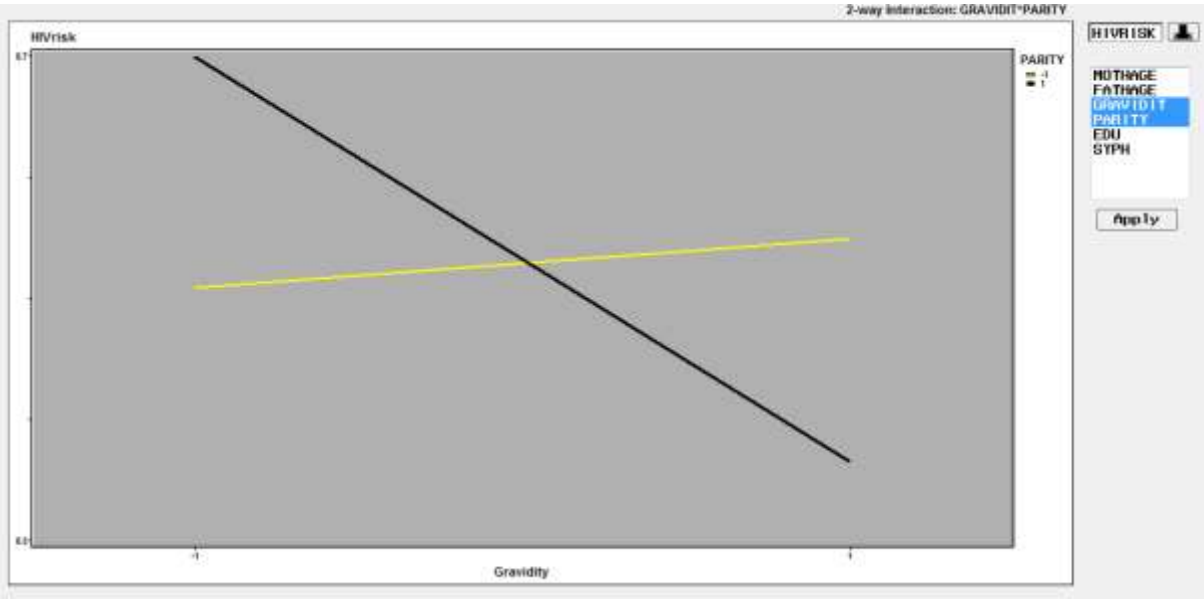
Interaction plot 2005: Mother's age * Partner's age



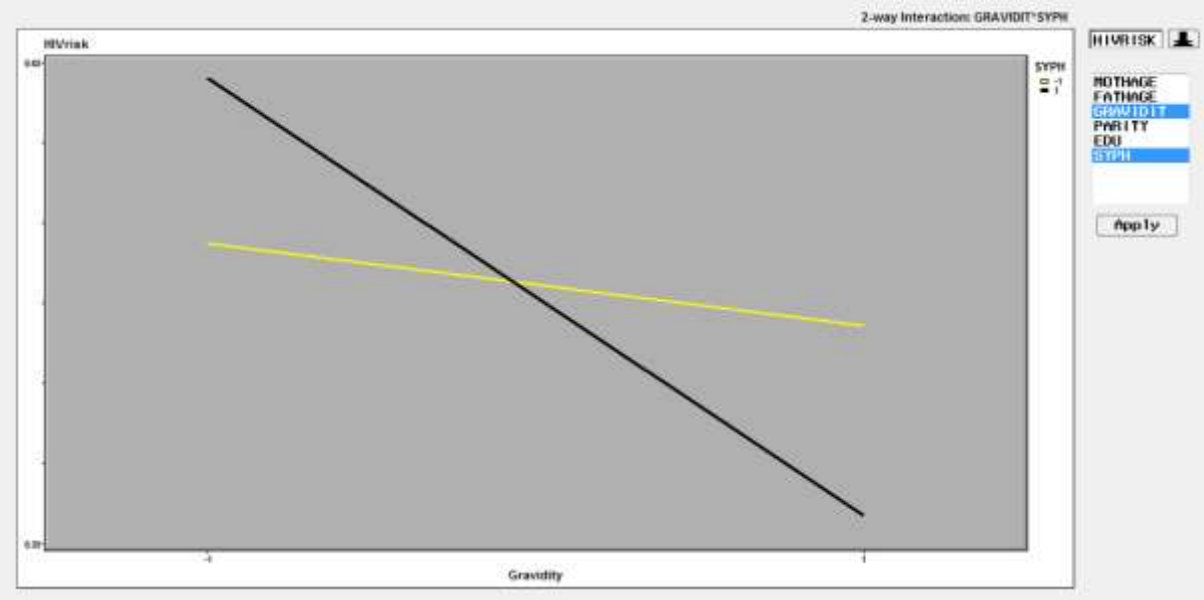
Interaction plot 2005: Partner's age * Education



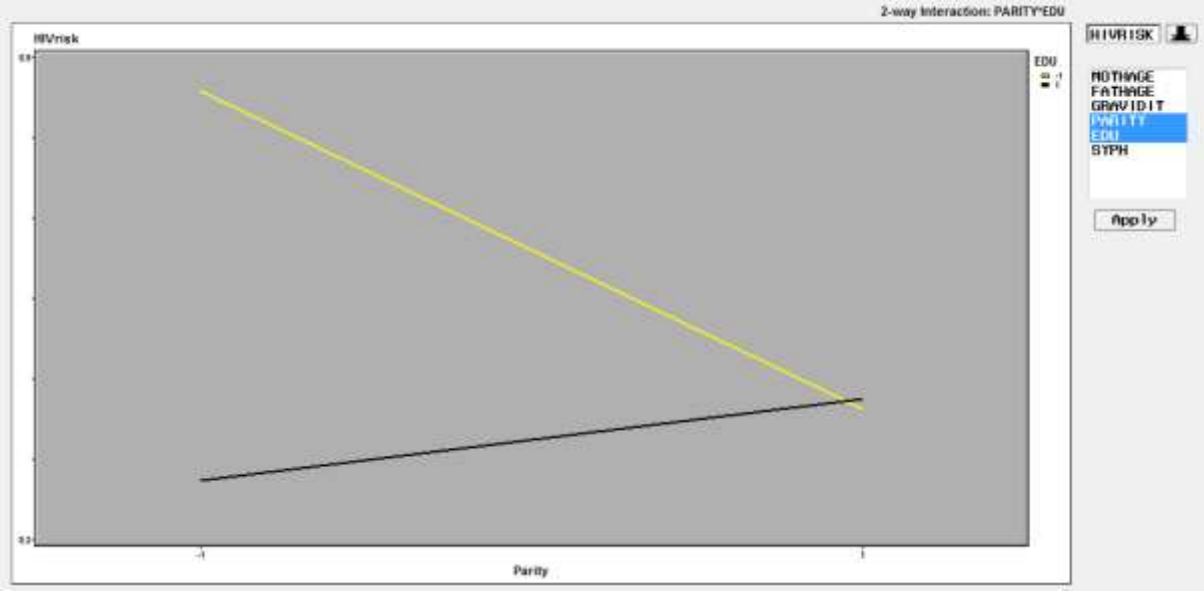
Interaction plot 2005: Gravidity * Parity



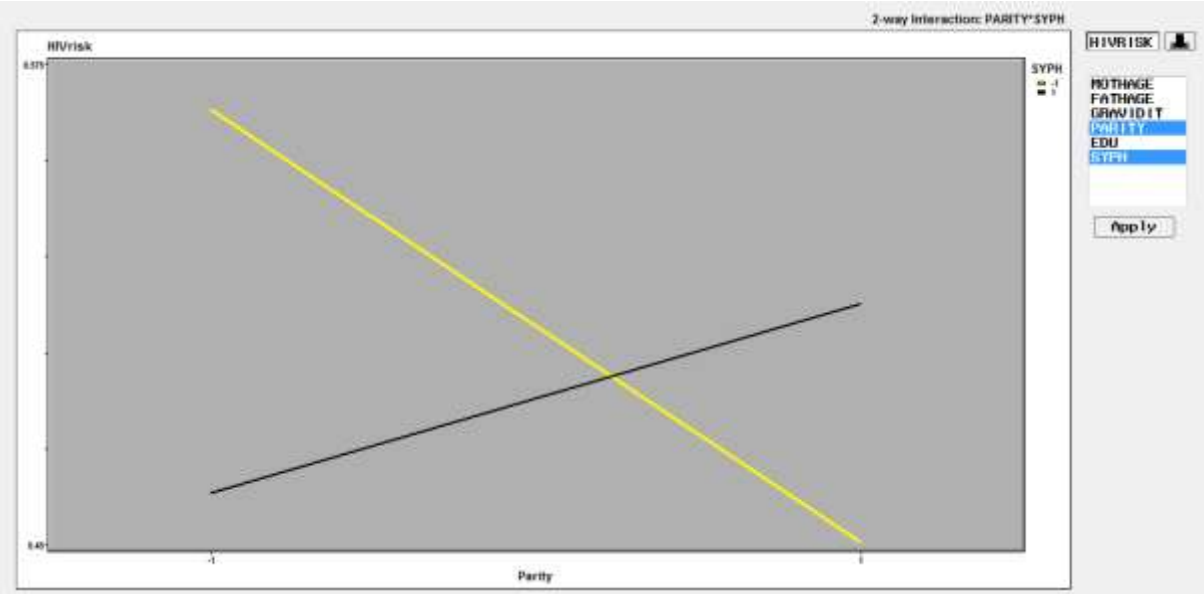
Interaction plot 2005: Gravidity * Syphilis



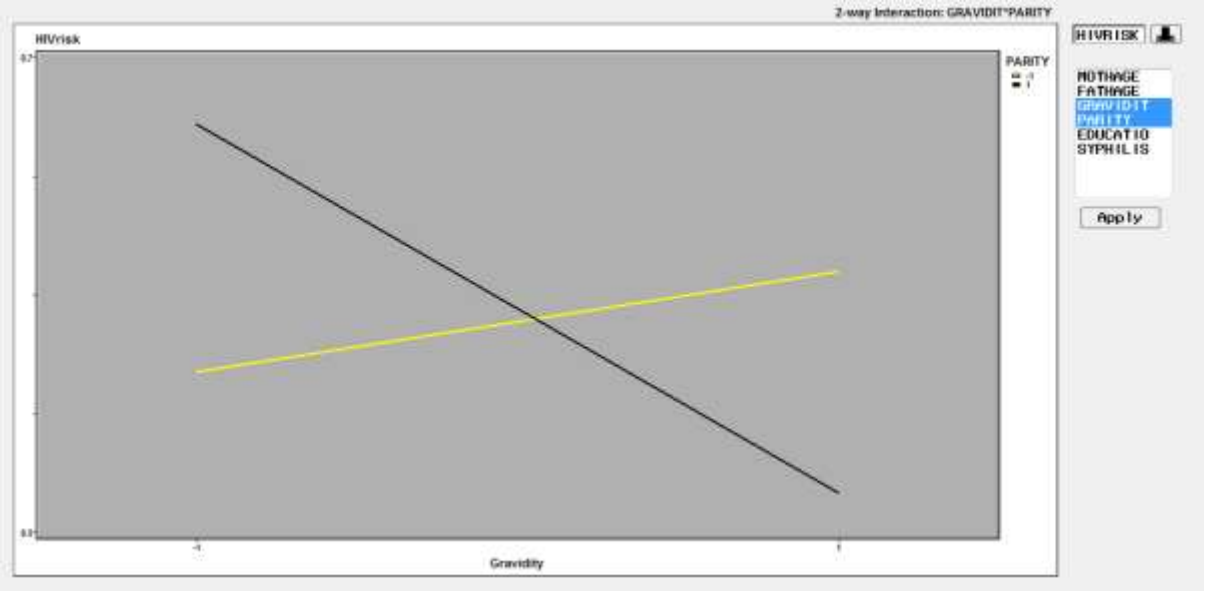
Interaction plot 2005: Parity * Education



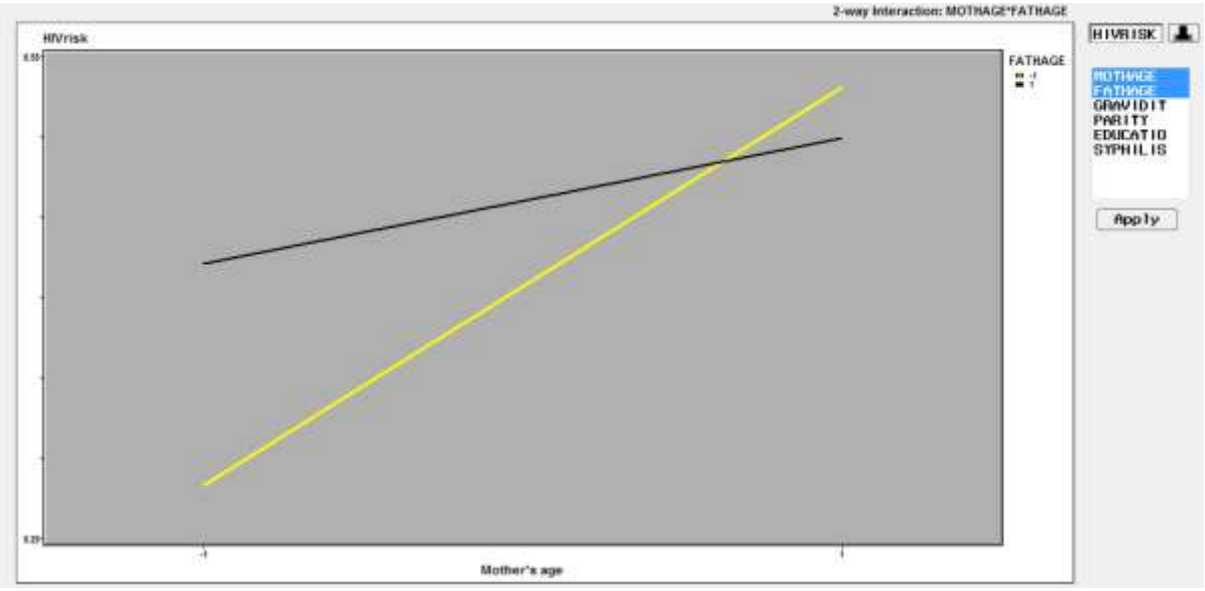
Interaction plot 2005: Parity * syphilis



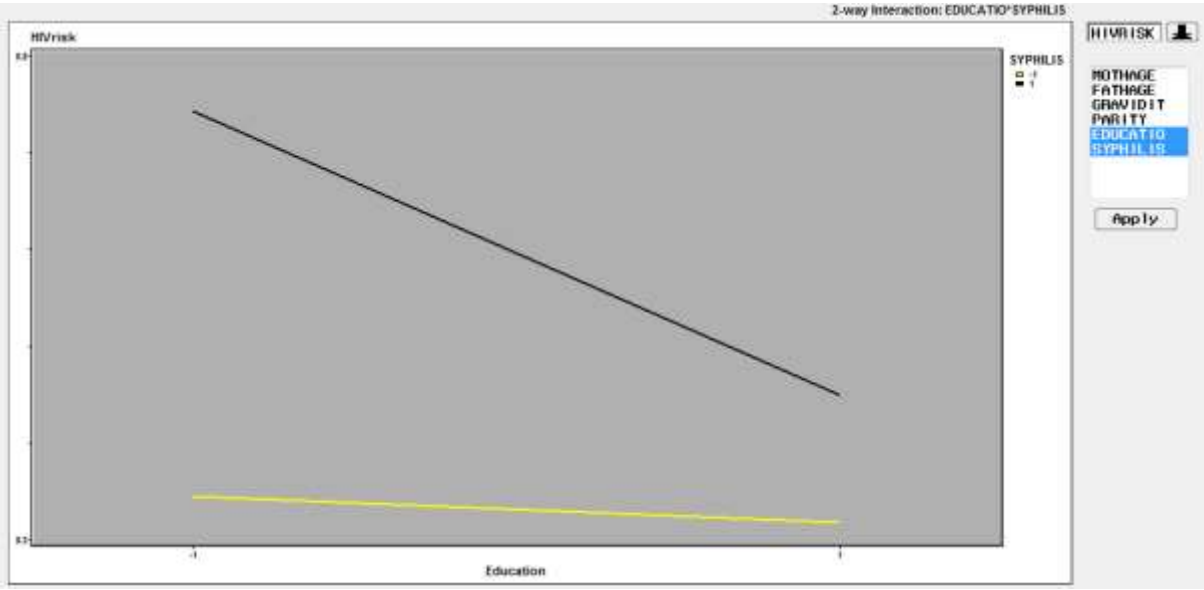
Interaction plot 2006: Gravidity * Parity



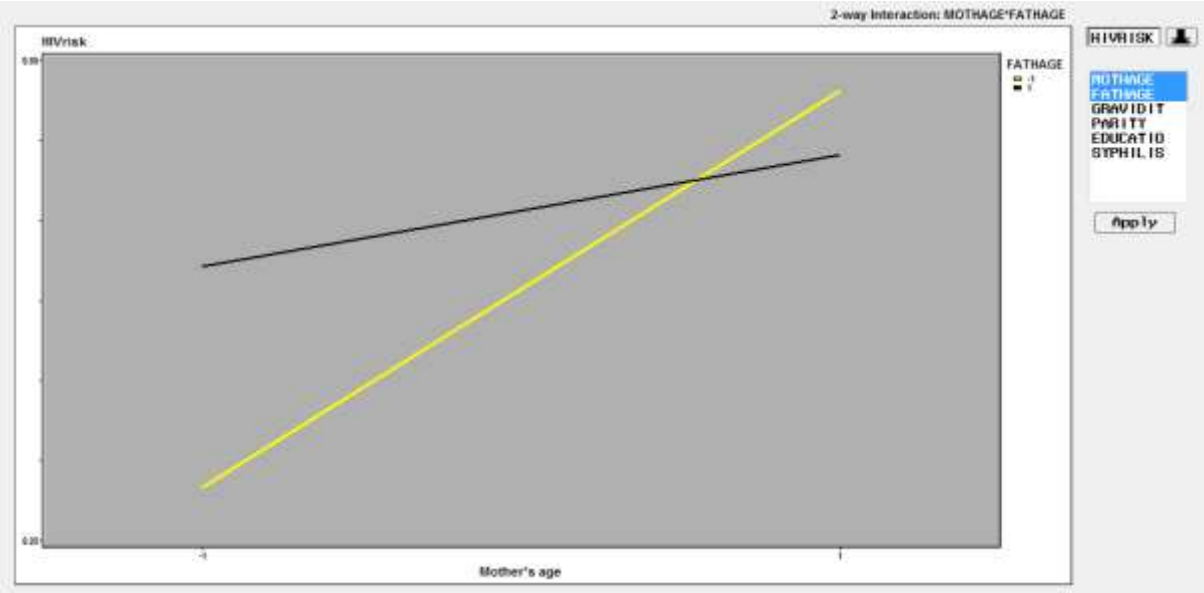
Interaction plot 2007: Mother's age * partner's age



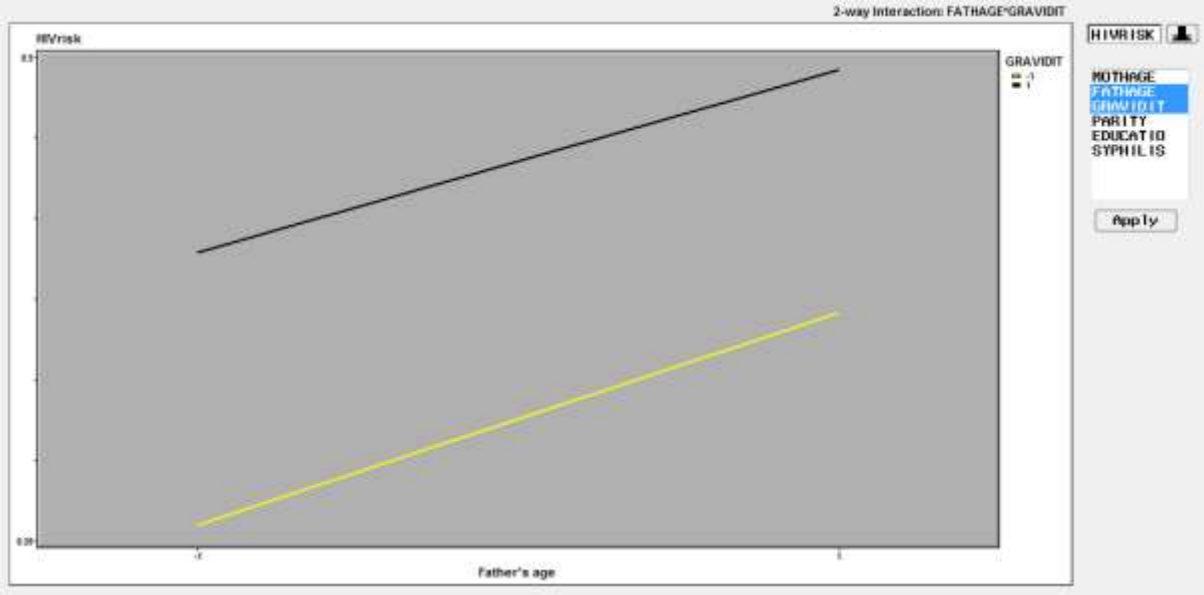
Interaction plot 2007: Education * syphilis



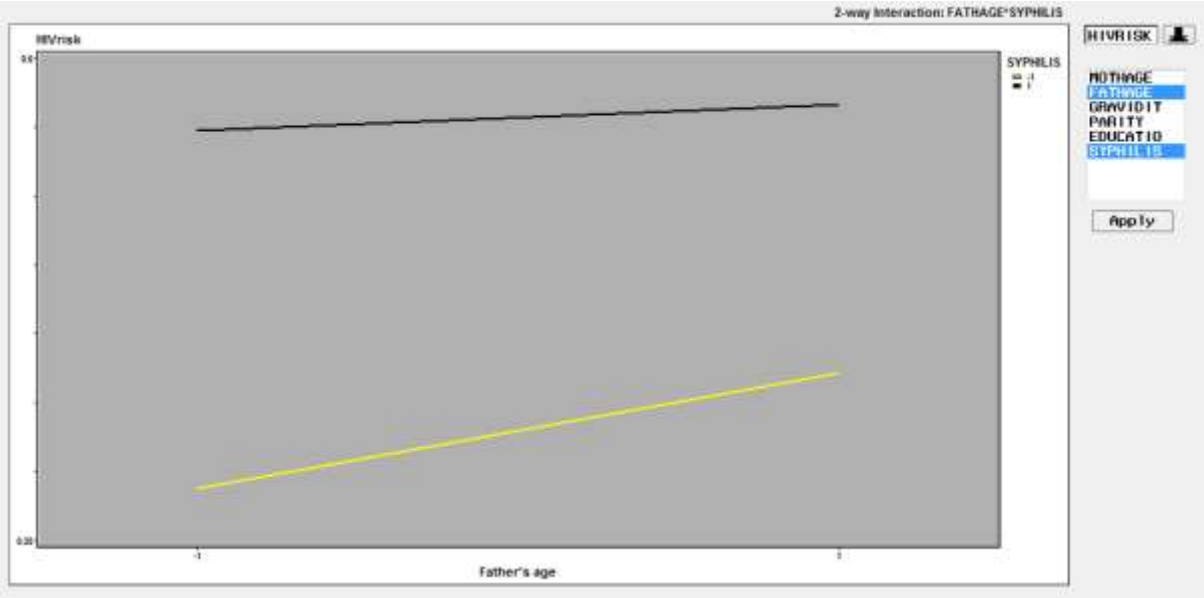
Interaction plot 2008: Mother's age * partner's age



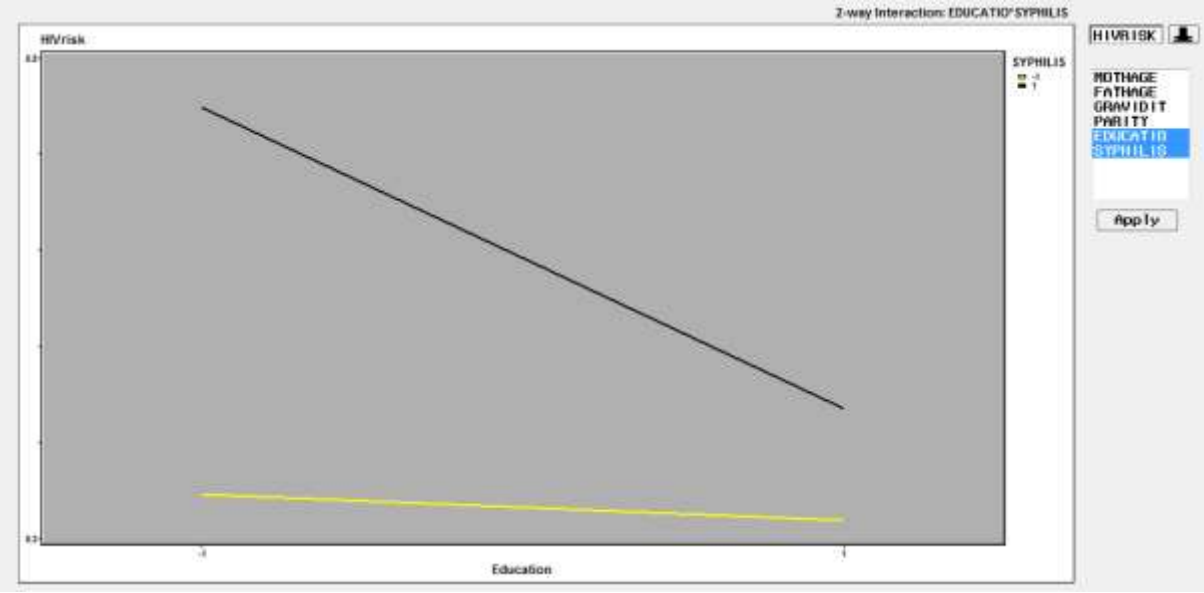
Interaction plot 2008: Partner's age * Gravidity



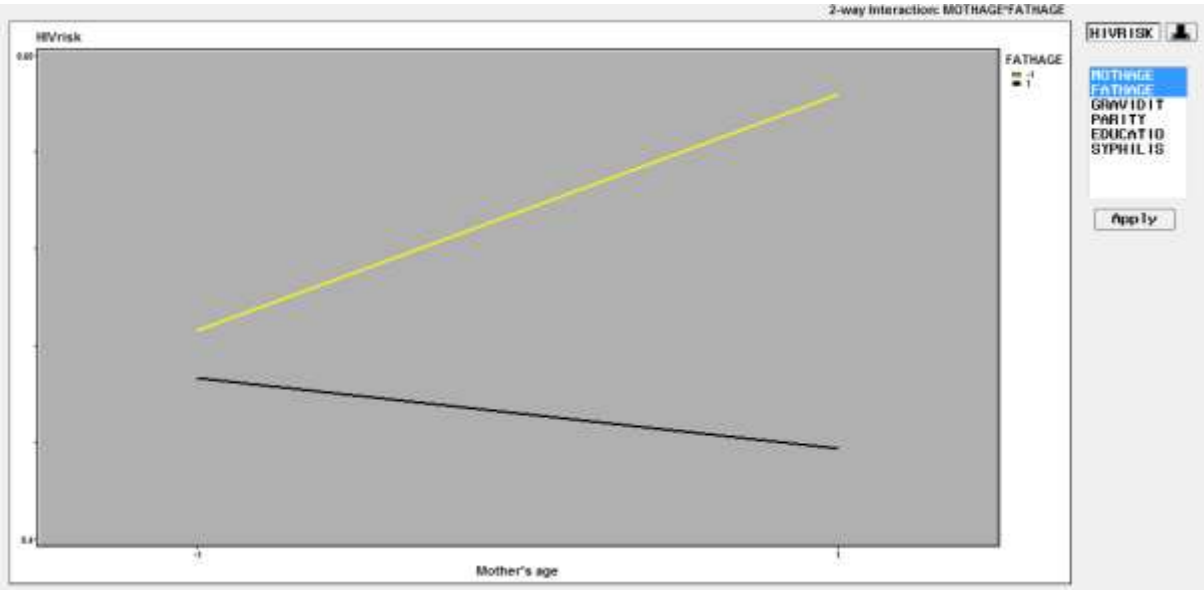
Interaction plot 2008: Partner's age * syphilis



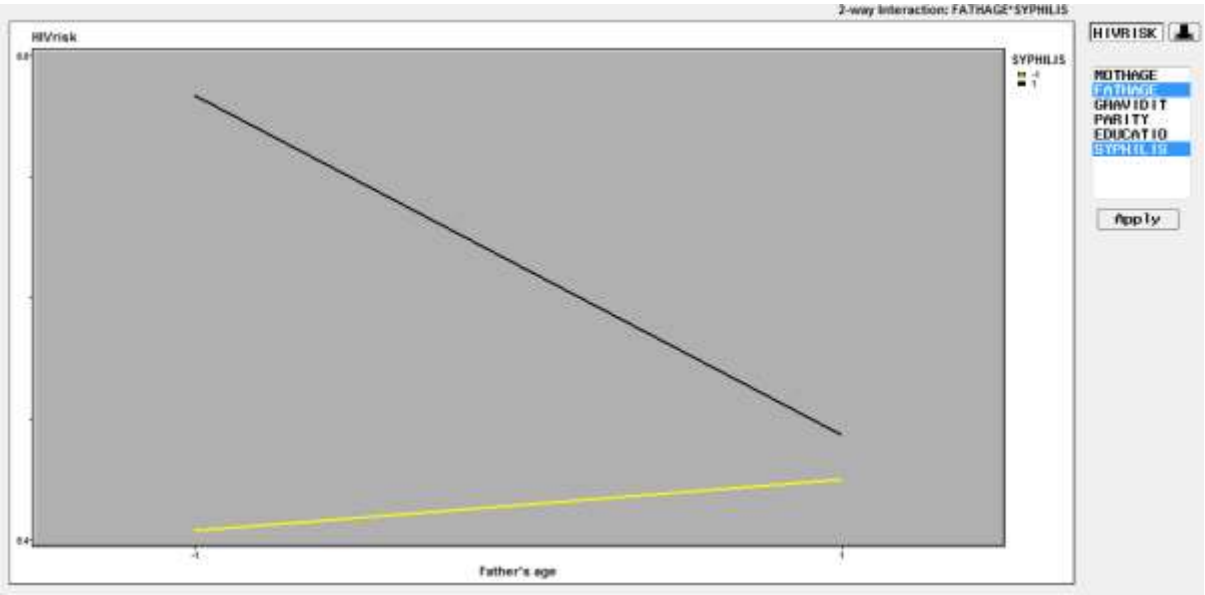
Interaction plot 2008: Education * Syphilis



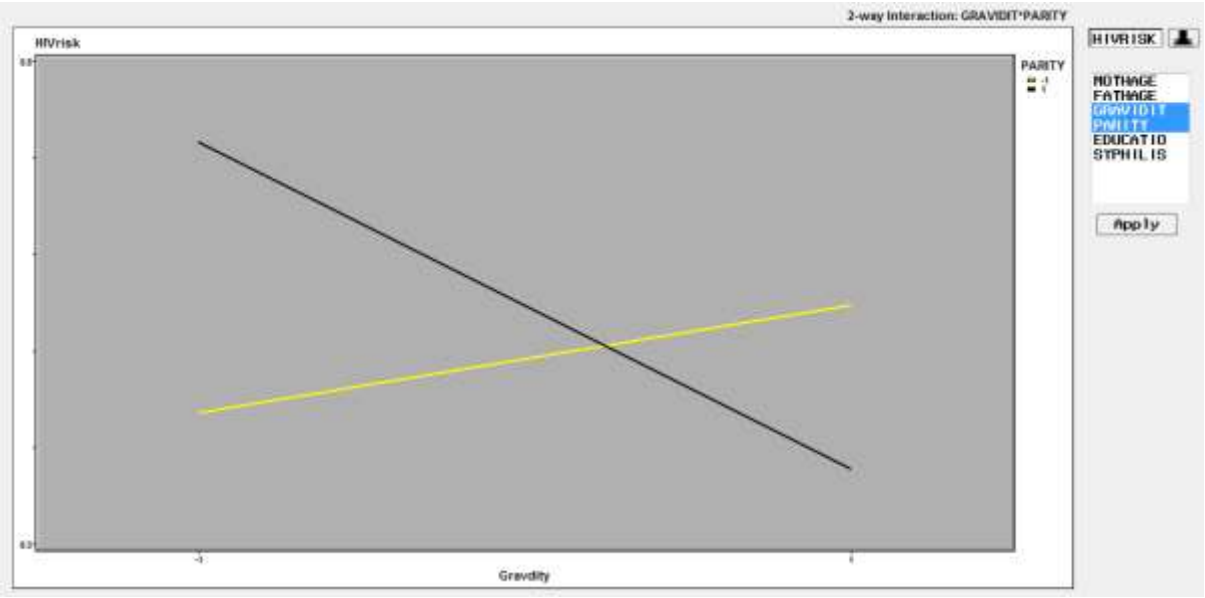
Interaction plot 2009: Mother's age * partner's age



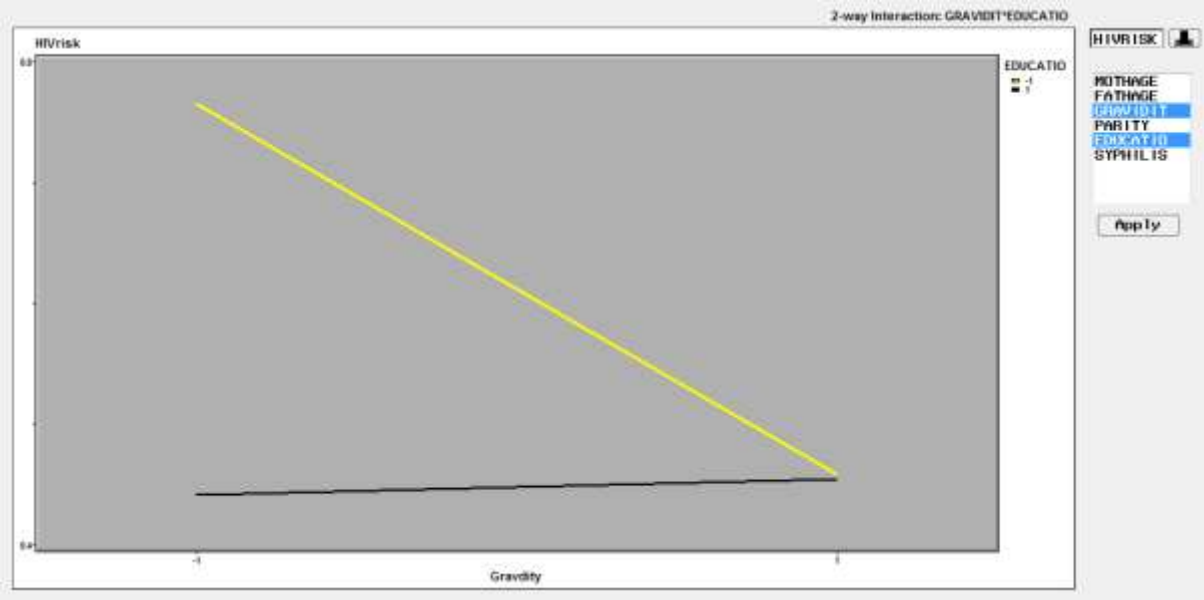
Interaction plot 2009: Partner's age * Syphilis



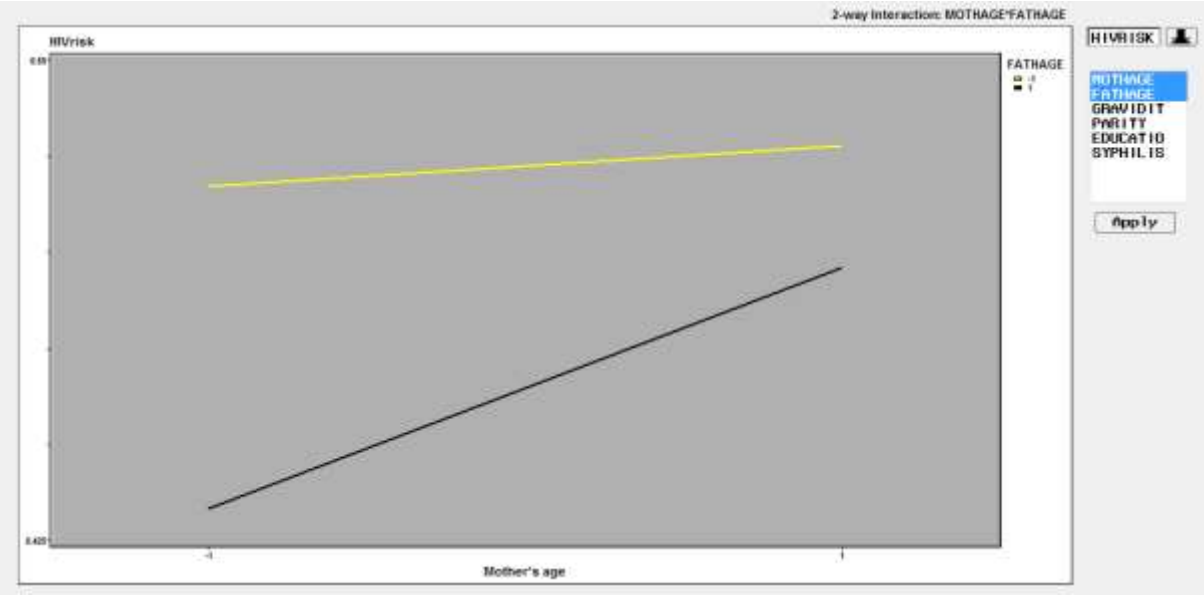
Interaction plot 2009: Gravidity * Parity



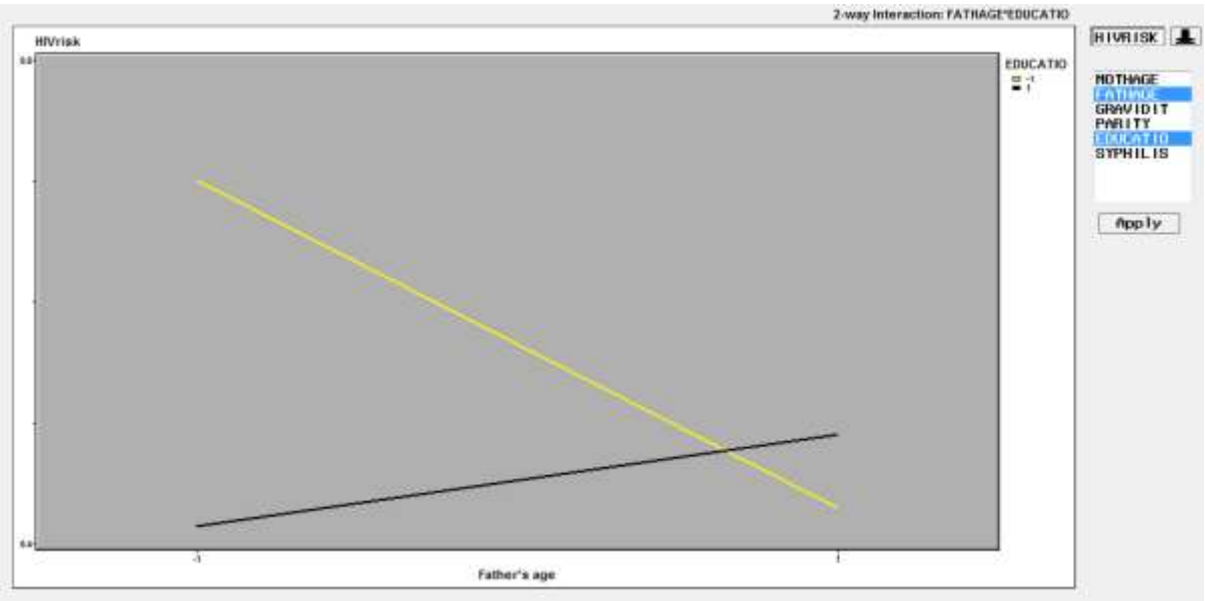
Interaction plot 2009: Gravidity * Education



Interaction plot 2010: Mother's age * Partner's age



Interaction plot 2010: Partner's age * Education



Interaction plot 2010: Gravidity * Parity